DOCUMENT RESUME

ED 093 915 TM 003 731

AUTHOR Huberty, Carl J.
TITLE Discriminant Analysis.
PUB DATE [Apr 74]
NOTE 85p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS *Classification; Computer Programs; Data Collection; *Discriminant Analysis; *Educational Research; Literature Reviews; Multiple Regression Analysis; Research Design; Sampling; *Statistical Analysis

ABSTRACT

Discriminant analysis is reviewed in terms of: (1) formulations, (2) interpretations, (3) uses, (4) issues and problems in applications, (5) recent developments and conceptualizations, and (6) general references and computer programs. Four aspects of a discriminant analysis are considered. They are: (1) separation: determining intergroup significant differences in terms of group centroids, (i.e., mean vectors), (2) discrimination: studying group separation with respect to dimensions and to (discriminator) variable contribution to separation, (3) estimation: obtaining estimates of interpopulation distances (between centroids) and of degree of relationship between the response variables and group membership, and (4) classification: setting up rules of assigning an individual to one of the predetermined exhaustive populations. Attention is given to the formal relationship between multiple regression analysis and two group discriminant analysis, as well as non-normal and incomplete data. Recommendations for reporting discriminant analysis results are given. A substantial list of references pertaining to discrimination and classification is supplied. (Author/RC)

ED 093915

TM 003 731

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

BEST COPY AVAILABLE

Discriminant Analysis

Carl J Huberty

University of Georgia

Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1974.

BEST COPY AVAILABLE

Discriminant Analysis

Introduction	1
Aspects of Discriminant Analysis	3
Separation	4
Discrimination	6
Linear Discriminant Functions	7
Requisite Data Conditions	9
Interpretation of LDFs	10
Variable Selection	15
Generalizability	21
Specific Uses of LDFs	22
Discrimination in Two-Group Case	23
Discrimination Research Applications	25
Estimation	28
Measures of Distance	28
Measures of Discriminatory Power	30
Classification	32
Requisite Information	32
Classification Rules	34
Efficiency of Classification	44
Estimating Error Rates	45
Classification in Two-Group Case	46
Specific Uses of Classification	49
Classification Research Applications	50
Other issues, Problems, Developments	53
Regression Analysis and Classification	53
Non-Normal Data	54
Incomplete Data	56
Use of T Versus E	57
Reporting Discriminant Analysis Results	58
General References	59
Computer Programs	60
References	63

Introduction

The purpose of this paper is to review discriminant analysis in terms of (1) formulations, (2) interpretations, (3) uses, (4) issues and problems in applications, (5) recent developments and conceptualizations, and (6) general references and computer programs.

A complete review of literature related to the development of various aspects of discriminant analysis will not be attempted in this paper. Excellent reviews have already been written which take the interested reader from the pre-Fisher conceptualization of the two-group classification problem, thru K-group formulations, to the use of discriminant analysis as a more general multivariate data analysis technique. The very comprehensive review by Hodges (1950), which focuses on the use of discriminant analysis for classification purposes, covers an historical development in the Pearsonian stage, dealing with measures of resemblance; the Fisherian stage, dealing with the linear discriminant function; the Neyman-Pearson stage, dealing with probabilities of misclassification; and the Waldian stage, dealing with risk and minimax ideas in classification. The review by Tatsuoka and Tiedeman (1954) covers developments in the area of classification as well as the relationship of discriminant analysis to other aspects of multivariate data analysis. In particular, they review C. R. Rao's conceptualization of the problem, extensions of R. A. Fisher's linear discriminant function, and the integration of the two. A brief review of early work in classification is provided by Ottman, Ferguson, and Kaufman (1956), along with an application of Rao's classification equations. A more recent review of classification theory and methodology is given by Das Gupta (1973). The review includes sections on the early history of classification

[basically a summary of Hodges' review], general classification problems and theory (including empirical Bayes approaches), multivariate normal classification, and non-normal distributions and nonparametric methods. An extensive and fairly recent bibliography is provided; references are listed for each section.

The formal relationship of the mathematics underlying the linear discriminant function -- "function" here is not used in a mathematical sense-- to other techniques in the domain of multivariate analysis was noted some years ago by, e.g., Bartlett (1947) and Tintner (1950), and more recently by Cooley and Lohnes (1971), Tatsuoka (1971), Van de Geer (1971), and Mulaik (1972). Despite the relationship, applications of "discriminant analysis" have in the past been somewhat divorced from other multivariate techniques, with classification being the primary concern. However, the use of discriminant analysis as an aid in characterizing group differences is seen as a very important extension from that as a mere classificatory tool. In his brief review, Tatsuoka (1969) states that the extension of discriminant analysis "...as a follow-up to MANOVA is probably one of the most significant developments in multivariate analysis during the past ten years. [p. 742]." Specific uses of discriminant analysis in relation to multivariate analysis of variance (MANOVA), and methods of interpreting (linear) discriminant functions are cogently reviewed by Tatsuoka (1973a).

In a somewhat restrictive view discriminant analysis had been considered in light of a mathematical problem. In this sense the idea was to simplify a multivariate situation to a univariate one. That is, given K well-defined groups and p measures on each individual in each group, the objective was to determine a (linear) composite of the p measures which would maximize the between-group variance of the composite relative to the within-group

variance. Once mathematical formulations of the basic problem were, in some ways, satisfactorily performed, applied statisticians and data analysts began to utilize them in various ways.

Aspects of Discriminant Analysis

In different areas of applications the term "discriminant analysis" has come to imply distinct meanings, uses, roles, etc. In the fields of learning, psychology, guidance, and others, it has been used for prediction (e.g., Alexakos, 1966; Chastian, 1969; Stahmann, 1969); in the study of classroom instruction it has been used as a variable reduction technique (e.g., Anderson, et al., 1969); and in various fields it has been used as an adjunct to MANOVA (e.g., Saupe, 1965; Spain and D'Costa, 1970). The term is now beginning to be interpreted as a unified approach in the solution of a reasearch problem involving a comparison of two or more populations characterized by multi-response data.

Discriminant analysis as a general research technique can be very useful in the investigation of various apsects of a multivariate research problem. In the early 1950's Tatsuoka and Tiedeman (1954) emphasized the multi-phasic character of discriminant analysis: "(a) the establishment of significant group-differences, (b) the study and 'explanation' of these differences, and finally (c) the utilization of multivariate information from the samples studied in classifying a future individual known to belong to one of the groups represented [p. 414]." Essentially these same three problems related to discriminatory analysis were mentioned some years later by Nunnally (1967, p. 388).

As a means of clarity in communication in this paper, four aspects of a "discriminant analysis" will be considered. They are (1) <u>separation</u> --

determining inter-group significant differences in terms of group centroids, (i.e., mean vectors), (2) discrimination -- studying group separation with respect to dimensions and to (discriminator) variable contribution to separation, (3) estimation -- obtaining estimates of inter-population distances (between centroids) and of degree of relationship between the response variables and group membership, and (4) classification -- setting up rules of assigning an individual to one of the pre-determined exhaustive populations. It should be noted that this terminology differs from that used by other writers. Of course, separation is usually thought of in terms of significance testing via MANOVA; in fact one-way MANOVA and "discriminant analysis" are sometimes considered synonymous (McCall, 1970, p. 1373). Discrimination as used here actually refers to methods of interpreting linear discriminant functions and their coefficients. This term has been used by others as the equivalent to what in this paper is called classification (Kendall, 1966, 1973; Kshirsagar, 1972). Rather than "classification," Rao (1965) uses "identification," while Kendall (1966, 1973) and Harman (1971) use "classification" as what is often referred to by behavioral scientists as "cluster analysis." The inclusion of estimation as an additional aspect was done for the purpose of emphasizing supplementary means of interpreting the results of a discriminant analysis.

Separation

The basics of MANOVA as a confirmatory (in the sense of significance testing) data analysis technique have been quite thoroughly covered in various books and technical papers and will not be discussed here. The formal equivalence, mathematically speaking, of MANOVA and some aspects of discriminant analysis was alluded to in the last section.

When the purpose of "research" is that of drawing conclusions and investigating scientific problems of group comparisons it has been suggested that "discriminant analysis" not be identified as a tool of educational research. Rather, it has been claimed that discriminant analysis applied to practical problems of optimal classification of individuals into groups. [See Bock (1966, p. 822).] In some investigatory situations, nevertheless, it may seem reasonable to use one-way MANOVA as a preliminary step to, or a first phase of, a discriminant analysis. The classical argument is that unless the investigator is assured of group differences to begin with, it is senseless to seek the linear composite to be used for (discrimination or) classification purposes. However, even though a value of any one of many possible MANOVA statistics might tend to support the null hypothesis of mean homogeneity, it is possible that for one reason or another the data support the alternative hypothesis.

If the mean differences among the criterion groups are all zero, no differentiation is of course possible in the normal case with equal dispersions, but it might be worth examining this special situation in the case of unequal dispersions. Bartlett and Please (1963) and Desu and Geisser (1973) cover ways of looking at this problem when there are only two groups.

In considering the role of MANOVA in a "discriminant analysis" the most important factor is the purpose of the analysis being performed and the questions one has of the data. The design of the study, including sampling, data collection, and questions ("contrasts" if you like), specify the data analysis technique(s). If the investigation entails some type of sampling of individuals with the notion of drawing conclusions, in an inferential sense, about levels of performance or about locations of

distributions, then MANOVA may be quite appropriate. This analysis may be followed up by what we have called discrimination and estimation methods. In this context, the variables whose means are being compared are the dependent variables while the independent variable(s) is(are) the grouping variable(s). On the other hand, the study may be one of prediction (of group membership), where the predictors are the independent variables and the dependent variable is a grouping variable. In this latter situation, there is no manipulation of the grouping variable, with the groups being formed a priori. Here, MANOVA may not be called for; the investigator proceeds directly to obtaining classification statistics.

Discrimination

A great deal of research in the behavioral sciences deals with the comparisons of different groups of individuals in terms of one or more measures. What characterizes a "group" depends, at least in part, on whether or not the grouping variable is manipulable -- as is experimental versus ex post facto studies. The MANOVA technique is often used in both of these situations when the data collection design is assumed to be appropriate. The omnibus null hypothesis tested in a one-way MANOVA design is that of the equality of the population centroids. When the populations are significantly separated, subsequent and more detailed study of the group differences would definitely be called for. One follow-up technique is that which we label as "discrimination"; others, such as multivariate multiple comparisons, are discussed elsewhere (Stevens, 1973; Tatsuoka, 1973a). Stevens (1972b) reviews four methods of analyzing between-group variation, one of which is based on linear discriminant functions.

Linear Discriminant Functions

The procedures used in discrimination center around linear discriminant functions (LDFs). The mathematics behind LDFs is presented in various books and papers (see, e.g., Tatsuoka, 1971; Porebski, 1966a). One resulting formulation may be briefly described as follows. A linear composite of measures on p random variables for individuals in K criterion groups,

[1] $$Y_1 = v_{11}X_1 + v_{12}X_2 + \ldots + v_{1p}X_p = \underline{v}_1' \underline{X} \ ,$$

is determined so that MSH_Y/MSE_Y is maximized or, equivalently, SSH_Y/SSE_Y is maximized. Here MSH_Y and MSE_Y denote the hypothesis and error mean squares with respect to Y-scores, respectively. (The "hypothesis" in a one-way MANOVA design refers to the between-group source of variation.) To obtain the v-values in [1], the largest non-zero characteristic root (or eigenvalue), λ_1, of $E^{-1}H$ is computed; i.e., the largest value of λ is obtained from the determinantal equation,

[1] $$|E^{-1}H - \lambda I| = 0.$$

The (pxp) matrices, E and H, are the error or within-groups and hypothesis or between-groups sums of squares and cross-products (SSCP) matrices, respectively. Then the (px1) eigenvector, $\underline{v}_1$, associated with λ_1 is found by solving the set of p equations,

[3] $$(E^{-1}H - \lambda_1 I)\ \underline{v}_1 = \underline{0}.$$

The elements of $\underline{v}_1$ are (within a constant of proportionality) the coefficients of the linear composite in [1]. As is well known, there may be more than one LDF. The succeeding roots, $\lambda_2 > \lambda_3 > \ldots > \lambda_s$ [where $s = \min(K-1,p)$],

yield discriminant functions that are mutually uncorrelated (in the total sample). The successive functions are determined so as to maximize relative separation after preceding functions are "partialled out."

In Van de Geer's (1971) integration of various multivariate techniques, the term "canonical discriminant factor analysis" is used to describe the process of extracting the LDFs. Harris (1954) and Pruzek (1971) use the term "dispersion analysis."

As we will see later, a means of interpreting LDFs is based on the number of functions to be considered. Here, as in interpretation of results in other domains of multivariate data analysis, parsimony is an objective. Data represented in a geometry of two-space, say, are more manageable and easier to interpret than if represented in spaces of higher dimensions. Thus, it behooves the researcher to discard discriminant functions which are judged not to contribute to group separation. This judgment can be subjective, in terms of the proportion of the total discriminatory power of p measures contained in a set of functions, or it can be based on statistical significance tests. The former judgment is based on ratios of individual eigenvalues to the sum of the eigenvalues. In the literature the process of testing the significance of a function has been lacking in clarity. First of all, Kendall (1968) has pointed out that such tests are "...not so much tests of the functions as tests of homogeneity (of population centroids) by the use of the functions. If heterogeneity is found, the function, ipso facto, is significant in the sense that it discriminates between real differences in an optimal way (except that we use estimators of dispersions and means instead of the unknown parent values) [p. 159]." Secondly, what hypothesis is of interest has not

been clearly stated in some writings. The issue on what hypothesis is being tested pertains to testing the significance of individual functions (or eigenvalues), or testing the significance of a set of functions after partialling out the complimentary set of functions that has earlier been judged to be significant. [The mechanics of both tests are given by Tatsuoka (1971, pp. 164-165). Two sources which leave the reader wondering which hypothesis is being tested are Eisenbeis and Avery (1972 pp. 63, 92-93) and Rulon, et al. (1967, p. 308). The test statistics reported in these latter two references are slightly in error -- N rather than N-1 is used in the test statistics, for one error.] The chi-square statistics used for these two hypotheses are different, but in a practical sense the conclusions are usually the same. That is, if it is concluded that the mth eigenvalue ($1 < m < s$) is the smallest one which is significant, then we usually will conclude that the last s-m eigenvalues (or functions) as a set with the first m removed do not yield significance (See Harris, 1974.)

Requisite Data Conditions

The validity of the generally used MANOVA tests of equal population mean vectors depends upon the conditions of multivariate normality and equal covariance structure being met (Bock and Haggard, 1968, pp. 110-113). The referrent distributions used for the various test statistics yield probability statements which may be somewhat distorted when either or both of the two conditions are not met. The degree and direction of distortion are not known. The multivariate analogue of the Behrens-Fisher problem (normality with unequal dispersions) is discussed by Anderson (1958, pp. 118-122). Ito (1969) has proposed alternative MANOVA tests to be used when either or both of the mentioned conditions

are violated; these tests show considerable promise for large samples. See also, James (1954).

Tests for assessing the fit of data to both multivariate normality and equal covariance are available. The tenability of the normality condition is typically assessed via goodness-of-fit tests, which call for large samples. Lockhart (1967) has proposed a partial testing procedure for the small sample case; more recent empirical studies have been made by Aitkin (1972) and Malkovich and Afifi (1973). A test of the equality of the group covariance matrices proposed by G.E.P. Box is presented by Cooley and Lohnes (1971, p. 229). The practical application of this latter test, and related concerns, are discussed by Porebski (1966b).

Interpretation of LDFs

We proceed, then, with our discussion of discrimination under the assumption that the two requisite conditions are, at least, tenable. Having established the dimensionality of the reduced space, it is of interest to give some interpretation of the, say, r "significant" LDFs. One very useful means of interpretation is provided by graphic methods. Even though the LDFs are mutually uncorrelated, they are not geometrically mutually orthogonal in the spaces of the predictor variables (Tatsuoka, 1971, pp. 163, 169). [In fact, the angle of separation between vectors representing two LDFs is an angle whose cosine is the inner product of the two corresponding (normalized) eigenvectors.] However, it is customary and convenient to graphically represent the K group centroids on the r LDFs by means of a rectangular coordinate system. The experience of this writer has shown that r is very seldom greater than two. That is, two LDFs generally

account for a great portion of the discriminatory power of the discriminators and, hence, a two-dimensional representation gives a fairly accurate picture of the configuation of the groups in the p-dimensional spaces. Of course, if r = 1, one can merely examine the numerical values of the K (pxl) mean vectors, $\bar{Y}$, to determine which groups or clusters of groups are separated from which other groups or clusters. If r = 2, a two-dimensional plot is helpful in interpreting the dimensions along which the K groups were found to differ. For example, consider the plot in Figure 1.

--Insert Figure 1 About Here--

From the graph it is clear that the first LDF discriminates Groups 2 and 4 from Groups 1, 3, and 5; whereas the second LDF discriminates Groups 1, 2, and 5 from Groups 3 and 4. If r > 2, pairwise two-dimensional plots may be used.

In making an interpretation of the resulting r LDFs, a substantive meaning of each function (or "canonical factor" or "canonical variate") is sometimes attempted. Two approaches have been employed. The first, in the sense of tradition, is based on magnitudes of function coefficients that are applicable to standardized scores. These "standardized weights" are found by multiplying each raw score coefficient by the within-groups standard deviation of the corresponding variable:

$$[4] \qquad v^{*}_{mj} = v_{mj} \sqrt{e_{jj}} \ ,$$

where e_{jj} is the jth diagonal element of E, m = 1,...,r, and j = 1,...,p. These weights have been considered by some writers (Tatsuoka, 1971, p. 170; McQuarrie and Grotelueschen, 1971) as if they were factor loadings.

Such a use of standardized weights as a means of interpreting LDFs has been criticized by Mulaik (1972, pp. 403, 422, 427) and Tatusoka (1973a, p. 280). This use of standardized weights may be questioned on theoretical grounds: these weights are actually partial coefficients and, hence, do not pertain to the common parts among the discriminators; two discriminators having large positive weights would not necessarily have anything in common which contributed to group separation.

The second approach that has been proposed for making a substantive or psychological interpretation of the LDFs is to use the correlations of the discriminators with the functions. The values of these correlations depend upon the data matrices used. The LDF coefficients may be obtained by using a "within-groups" formulation as reflected in equation [3], or via a "total-group" formulation which, in essence, is a canonical correlation attack on the problem (Tatsuoka, 1971, p. 177). The matrix product used to get the (pxs) total-group (canonical) structure matrix is simply

$$[5] \qquad S_T = RV \ ,$$

where R is the (pxp) correlation matrix based on T (= E + H), and V is the (pxs) matrix consisting of the s LDF coefficient vectors. The structure matrix containing the within-groups correlations is given by

$$[6] \qquad S_W = D^{-1} \, C \, V^* \ ,$$

where D^2 = [diag C], C is the (pxp) within-groups covariance matrix [= E/(N-K)], and V* is the (pxs) matrix of s LDF standardized weight vectors. As might be expected, the correlations determined by [5] will be larger than corresponding correlations in [6]. In terms of labeling

the functions, the resulting interpretations based on [5] or [6] will be the same. Such interpretations are, at best, a very crude approximation to any identifiable psychological dimensions.

Darlington, Weinberg and Walberg (1973) contend that the choice between standardized weights and correlations for interpreting LDFs ought to be based on the practical consideration of sampling error; they argue that because of greater stability, correlations ought to be emphasized, at least in some cases. In a Monte Carlo study, Huberty and Blommers (1972) concluded that neither statistic, when based on the leading LDF, was very stable in a cross-validation sense; this conclusion was not fully supported by Thorndike and Weiss (1973). More will be said on this in the section, "Generalizability."

There has been some attempt to achieve greater interpretability by rotating LDFs. Tatsuoka (1973a, pp. 301-302) briefly reviews two studies in which rotation was used; the matrix to be rotated in one study (Anderson, Walberg, and Welch, 1969) consisted of the (total-group) variable - LDF correlations, and in a second (McQuarrie and Grotelueschen, 1971) consisted of standardized weights. It is questioned by the present writer whether or not rotation of such canonical factors will, in most situations, be of great help in interpretation. Tatsuoka (1971) states that rotation (of a structure matrix, at least) "...requires further scrutiny and theoretical justification...[p. 301]." The issue of oblique versus orthogonal rotation of LDFs is a theoretical one yet to be resolved. Two general methods of rotation are discussed by Hall (1969); one method which attempts to arrive at an interpretable taxonomy of variables involves an orthonormal rotation of a structure matrix.

This discussion of LDF interpretation may be closed with a caveat: unless an investigation deals with variables having some common psychological grounds, attempting to substantively interpret the LDF(s) may be wasted effort. Thus if substantive interpretation of the function(s) is important to an investigator, his initial choice of variables ought to be made carefully.

The problem of variable contribution to group separation in discriminant analysis is a sticky one, as it is in multiple regression analysis (Darlington, 1968). It may be argued that the variables act in concert and cannot logically be separated. As far as an index to measure the "importance" or the size of the "effect" of a variable is concerned, no completely satisfactory proposal has been made. Traditionally the statistic used to assess the contribution of each variable (in the company of all others) has been its standardized weight. The variable - LDF correlations discussed previously in terms of substantive interpretation have also been suggested to order variables in terms of their contributions to separation. In a discussion which only involved the leading LDF, Bargmann (1970) argues that if correlations for only some of the variables are large (in absolute value), and small for others, then the former variables contribute essentially to group separation. If the order in which variables are entered into the analysis can be determined a priori, then the step-down procedure of Roy and Bargmann (1958), discussed by Bock and Haggard (1968) and Stevens (1973), can be used for testing the significance of the contribution of each newly entered variable.

This section is closed with a proposal for a procedure of analyzing data for the purpose of discrimination involving K (>2) groups. This "hierarchical analysis," which is only appropriate when the grouping variable

is clearly categorical as opposed to being ordinal, may be described as follows. First of all, the variables are screened (as discussed in the next section); suppose that after this screening there are p variables remaining. As in equation [3], the eigenvector, $\underline{v}_1$, associated with the largest eigenvalue of $E^{-1}H$ is determined. The correlations between each discriminator and the linear composite of all p variables, $\underline{v}_1' \underline{X}$, are then found -- the first column of S_w in [6]. An argument for using only the first LDF is presented by Bargmann (1969, p. 573). The variables are ordered on the basis of the absolute values of these correlations. The ordered array is then examined for a "breaking point" (or possibly several, if p is large) between large and small absolute values. If a disjunction occurs, then the variables fall into two classes with respect to discrimination. New (leading) LDFs for the two subsets of variables are calculated, and the process is repeated until no new subsets can be generated. (At each step variable - LDF correlations are determined and a substantive interpretation may be attempted.) A hierarchy of sets of variables, based on directly observable and, hence, interpretable measurements can be thus established. This, it may be argued, is preferable to an interpretation of residual discrimination -- that associated with $\underline{v}_2$, say, after the elimination of an artificial variable.

Variable Selection

The process of selecting variables in discriminant analysis, as in any multivariate analysis, can be considered before or after the main analysis. If Cochran's (1964) conclusions can be extended from the two-group to the K-group case, the operation of discarding noncontributing discriminators at the outset may be hazardous. However, many statisticians

suggest that unless a variable is "significant" in a univariate sense, it is probably wasteful to include it in a multivariate analysis, even if it correlates appreciably with good discriminators. Grizzle (1970) recommends that variables which do not have a reasonable expectation of containing information about group differences should not be included in the analysis; this would prevent a loss of power. The argument presented is based on the idea that the deletion of a non-significant variable does not change the largest characteristic root (λ_1 from equation [2]) very much. To conclude, preliminary to data collection variables ought to be chosen judiciously, on the basis of theory and prior research (Tatsuoka, 1969, p. 743). Then, following collection of data on the p chosen variables, p univariate analyses are performed; those variables not yielding significance at a low probability level are deleted prior to the multivariate analysis. A possibly extreme situation is as follows. Assuming univariate ANOVAs are appropriate, clearly if the "signal-to-noise" ratio (F-value) for a variable is less than unity, eliminating the discriminator from further consideration is the sensible thing to do.

The problem of variable selection or deletion may also be of interest after the initial multivariate analysis has been carried out. In many situations involving discrimination the investigator is presented with more discriminator variables than he would like and there arises the question of whether they are all necessary and, if not, which of them can be discarded. That is, having obtained the linear composite, the investigator may ask if the data might not have been adequately explained by using a subset of the original p discriminators. The objective is to include as many variables as possible so that reliable results may be obtained, and

yet as few as possible so as to keep the costs of acquiring data at a minimum. Reasons for reducing the number of discriminators may be summarized as follows [see Horst (1941) for elaboration]: (1) to obtain fundamental and generally applicable variables, (2) to avoid prohibitive labor, and (3) to increase the sampling stability of the LDF(s). On the last reason Horst mentions that as the ratio of the number of discriminators to the number of individuals increases, "...there is a tendency for the accuracy of (discrimination) to decrease if the weights determined on the first sample are applied to a second group [p. 102]."

There is a dearth of literature covering the problem of variable selection or reduction in multiple-group discriminant analysis. No reasonably optimal procedure has yet been developed for discarding variables; reasonable in the sense of amount of calculation, and optimum in the sense that the selected variables would yield the maximum amount of separation among the groups for that number of variables. Of course, one could consider all possible subsets of the original p variables, but, just as in multiple regression analysis, this is very expensive. Six "selection" procedures were reviewed by Huberty (1971a). The objective of one procedure is to obtain a subset of variables that may be considered representative of the complete set (Bargmann, 1962a). Representativeness is based on a (maximum likelihood) factor analysis of the discriminator within-groups intercorrelation matrix, followed by an oblique rotation of the resulting factors. The usual eigenanalysis (see equation [3]) is then performed on the variables that define each factor. Correlations between each of these variables and the first (or leading) LDF are determined; i.e., the first column of S_w in equation [6]. Variables are then selected that load on

each factor and correlate highly with the respective leading LDF. Another procedure, suggested by Horst (1965, p. 555), involves a principal component analysis of the discriminator within-groups intercorrelation matrix, followed by an orthogonal rotation of the resulting components. A subset of variables is selected such that each component will be adequately represented in the subset. Variables are selected which have the highest loadings (in absolute value) on each of the components; persumably, no variables are selected which have high loadings on more than one component. The other four procedures discussed yield some type of ordering of the predictors; they are based on (1) standardized weights for the first LDF, (2) univariate F-ratios, (3) discriminator versus (first) LDF correlations, and (4) an ordering provided by the BMD 7M stepwise program (Dixon, 1973).

An empirical comparison of the six procedures using two sets of data -- $K = 3$ and $p = 13$ for one set and $K = 5$, $p = 17$ for the other -- was made using the criterion of the proportion of correct reclassifications of individuals across all procedures for subsets of a given size. It was found that the stepwise procedure yielded the best subsets in the sense of most accurate classification. It should be pointed out that an objective of the procedures involving a dimension (i.e., factor or component) analysis is to select a subset that is representative of the total set. Selecting a representative subset and one that will have nearly the same discriminatory power as the original set will not necessarily characterize the subset selected simultaneously.

As suggested in the preceding section of the present paper, standardized weights associated with a given LDF (first or otherwise) may be used to assess the contribution of each variable (in the company of all others) to the separation accomplished by that LDF. This method of assessment may be extended to obtain a measure of the relative contribution of each

variable to the total set of LDFs. The set of weights are weighted by a function of the proportion of discriminatory power accounted for by each LDF. The measures to be used are given by the pxl vector

[7] $$\underline{a} = V^* \underline{\lambda} \quad ,$$

where V* is the pxs matrix defined in [6], and $\underline{\lambda}$ is the sxl vector of eigenvalues of $E^{-1} H$. Another proposal is made for ordering variables, with respect to group separation, on the basis of the variable versus LDF correlations. The measures to be considered are somewhat analogous to "communalities" in factor analysis. These are given in the pxl vector,

[8] $$\underline{b} = [\text{diag } SS'] \quad ,$$

where S is the pxs structure matrix defined in [5] or [6]. The jth element in $\underline{b}$ is the sum of squares of the "loadings" in the jth row of S.

Various selection schemes need to be researched further. The need exists for empirical studies of other stepwise procedures; e.g., that proposed by Dempster (1963), which is a forward stepwise procedure with the variable ordering determined by a principal component analysis. Hall (1967) has proposed a forward procedure involving multivariate analysis of covariance (MANCOVA) which, in essence is the same as that used in the BMD stepwise program; the variables already in the analysis are the variates while the remaining variables are the covariates. A variation of this use of MANCOVA to select the most effective discriminators is given in a study by Horton, Russell, and Moore (1968). See also, Smith, et al. (1972). Hotelling's trace statistic was used as a criterion for selecting variables in a forward manner by Miller (1962). There is some argument for using a "backward" scheme, where

variables are deleted from, rather than added to, the analysis [see Mantel (1970)]. It would also be of interest to study the appropriateness of the measures in [7] and/or [8] as indicators of variable contribution.

Eisenbeis and Avery (1972) suggest a variable selection method which incorporates a number of techniques used jointly. Such a combination method may be described as follows. Determine variable orderings based on a number of different techniques -- e.g., stepwise, standardized weights, variable-LDF correlations, backward elimination. To determine an upper bound on the number of variables to be retained, a minimum arbitrary acceptance level of the reduction in discriminatory power using a set of size q instead of all p variables is set. Eisenbeis and Avery support the one percent significance level of a MANCOVA F-statistic as a criterion. Other criteria such as a significance level of Hotelling's trace statistic, or a specified proportion of correct classifications yielded by the set of entered variables may also be used. This may give t (the number of techniques used) different subsets of size q. Two different approaches may now be taken to arrive at a single subset. One approach, mentioned by Eisenbeis and Avery (1972, p. 82), is to determine those of the q variables and of the p-q variables that are common across the different techniques. The latter variables, say m in number, are discarded from further consideration, and the former, say n in number, are to be included in the final subset of size q. To obtain the final subset of size q, then, the best q-n are selected out of the p - (m+n) questionable variables by considering all possible subsets of size q-n. A second approach, similar to that suggested by Draper and Smith (1966, p. 172) for use in multiple regression, is to consider all possible subsets of size q from the p original variables.

Finally, it is noted that after a subset of variables has been selected it is desirable to reanalyze the data only on the selected

variables so as to assess their relative contribution. This is particularly true when the assessment is based on standardized weights; the rank-order of the selected variables as a set by themselves may be different from their rank-order when considered in the company of all the original discriminators.

Generalizability

Even though discrimination involves basically exploratory techniques, very often its users attempt to generalize results to other sets of subjects, other variables, or other situations. Generalizability may be thought of in terms of statements of inferences from sample results to some population, and in terms of stability of the obtained results over repeated sampling. Mulaik (1972) emphasizes the caution with which one proceeds in making inferences when treating LDFs as factors. One warning is that with the formulation of [3], the LDFs obtained pertain to the variables after variance in them due to group differences has been removed from them. Thus, such dimensions do not reflect variance which exists in the variables on which the groups differ and "...may in some contexts give misleading characterization of the nature of the discriminant functions [p. 428]."

Not much conclusive evidence has been found regarding the stability of results in discrimination studies. In a Monte Carlo investigation designed to study the comparative stability of standardized weights and variable-LDF correlations, Huberty and Blommers (1972) found that neither index held up to any great extent under repeated sampling. (It should be noted that only the leading LDF was considered in that study.) Two sets of live data were used in a study by Thorndike and Weiss (1973) who concluded that if an investigator "uses a single sample and attempts

to interpret the canonical components (LDFs) ..., he may be interpreting nothing more than sample-specific covariation [p. 131]." They did conclude, however, that component loadings (variable-LDF correlations) are consistent in cross-validation and in this sense are more stable and more useful than standardized weights. Stevens and Barcikowski (1974) concluded from a Monte Carlo study that in some situations (depending upon variable intercorrelations) standardized weights are more stable than variable-canonical variate correlations, and in other situations the reverse holds.

Problems of generalizability due to instability of some results appear to point to the need for replication of studies and cross-validation of findings. Of course, the use of simple (or double) cross-validation techniques call for relatively large samples. [Tatsuoka (1970, p. 38) suggests that in a usual discriminant analysis the size of the smallest group be no less than the number of variables used, p. This may be a bit conservative.] To use cross-validation techniques it is recommended that the smallest n-value be at least as large as 3p. Then in the cross-validation process, a random one-third of the total number of observations may be withheld from each group to serve as a "holdout sample." Horst (1966) points out the dilemma into which one is placed when using cross-validation techniques: "If we develop a procedure and then cross-validate it, we have ipso facto not developed the best procedure possible from the available data [p. 140]."

Specific Uses of LDFs

The use of LDFs as an aid in the interpretation of MANOVA results was mentioned in the "Introduction" of this paper as a major breakthrough in multivariate analysis. Uses of LDFs in factorial MANOVA are illustrated

by Jones (1966) and Saupe (1955) who point out that multiple LDFs are useful in interpreting the source of significant interaction effects as well as the source of significant main-effect differences. Both of these writers base their interpretations of LDFs on standardized weights in preference to variable-LDF correlations. Some writers (e.g., Timm, 1974) prefer the use of simultaneous test procedures (Gabriel, 1968) for studying significant differences, while others (e.g., Tatsuoka, 1973a) prefer the LDF approach.

Tatsuoka (1973a, p. 284) also suggests that LDFs may be helpful in deciding when to terminate a clustering procedure such as that of Ward (1963). At each stage of the analysis the LDFs based on the clusters (of individuals) determined to that point can be examined for interpretability. Discrimination procedures were used by Rock, Baird, and Linn (1972) as a follow-up to a cluster analysis involving areas of study of college students. In addition to univariate F-values, discriminator-LDF correlations were considered in assessing relative contribution of the variables to the obtained first LDF.

Techniques of discrimination have been shown to be of help in the study of pattern recognition. The research of Kundert (1972) illustrates the use of an LDF in assigning scale values to categories of a response variable, irrespective of the manner in which the categories may be ordered.

Discrimination in Two-Group Case

The relationship between multiple-group discriminant analysis and canonical correlation was pointed out previously. The lower level relationship between two-group discriminant analysis and multiple correlation has been the subject of many writings. The proportionality of the raw score coefficients for the two analyses was shown by Michael and Perry (1956). More recently this proof has been vastly simplified via the use of matrix

notation (Healy, 1965; Porebski, 1966a; Cramer, 1967; Tatsuoka, 1971, pp. 171-173). Because of this relationship, many of the methods used in interpreting a regression analysis are applicable in two-group discriminant analysis. Collier (1963) showed that tests used in deleting variables in regression analysis and in discriminant analysis are equivalent, while Huberty (1972b) showed that predictor variables may be equivalently ordered (with respect to contribution to separation) by univariate F-ratios and by within-groups variable versus LDF correlations. Cochran (1964), Weiner and Dunn (1966), and Urbakh (1971) have also studied the problem of eliminating variables in the two-group case.

In studying discrimination between two groups of foreign graduate students in business administration, Grimsley and Summers (1965) applied tests of significance of the LDF coefficients -- see Kendall (1968, p. 163)-- in determining the most effective combination of discriminators to differentiate between success and failure groups. Recently, Eisenbeis, Gilbert, and Avery (1973) studied the problem of variable assessment and selection in the context of a specific empirical problem. They concluded that the various selection methods studied "...could yield radically different inferences about the relative power of individual variables [p. 218]." It was also concluded that "... the assessment of the relative performance of the different subsets also varies depending upon whether the goal is to select the subset that maximizes differences between group means or to choose the combination of variables that yields the best classification results [p. 218]." Huberty (1974b) has shown the formal equivalence between a test for deleting variables which is based on distance and a test based on MANCOVA for the two-group case. Implications of this equivalence for interpretation of results of multi-group analyses were

were discussed in the section, "Variable Selection."

Discrimination Research Applications

No attempt will be made to review all studies in behavior research that incorporate discrimination procedures. Rather, selected journal articles will be cited so as to acquaint the reader with (1) some research situations -- i.e., types of subjects, criterion groups, and discriminators -- for which discrimination may be helpful, and (2) the discrimination techniques being used. No critique of the substantive discussions and conclusions presented in the articles will be attempted. The articles reviewed are in addition to those discussed earlier in this paper and all appeared in 1968 or later. Huberty (1969) cites 30 studies reported from 1963 to 1968 in which discriminant analysis techniques were used.

The entire focus of one study was a two-group simple MANOVA, although the analysis technique was described as a "multiple discriminant analysis." Williams (1972) used six factor scores on a semantic differential for differing socioeconomic status groups -- low versus middle -- of 181 fifth grade inner city school children. Another study (Maw and Magoon, 1971) involved a 2x2 MANOVA design with sex and curiosity as the grouping variables. Twenty-six measures on affective, cognitive, personality, and social trait variables were obtained on the four groups of middle class white fifth-graders. Since sex-by-curiosity interaction was not significant, the associated LDF was not considered for interpretation. The variable-LDF correlations (the type was not specified explicitly) were used in interpreting the sex and the curiosity LDFs. Canonical correlations for sex and curiosity, each versus the 26 variable composite, were also examined.

Two studies were found in which the LDF interpretation was based on standardized weights. Project TALENT data were used by Schoenfeldt (1968) in a study involving a random selection of about 300 students in each of six post high school education groups. Measures on 79 variables were

available; after preliminary screening, 26 were selected for use. A 64-item study strategy questionnaire was used by Goldman and Warren (1973) to obtain data on 538 university students who were in four different undergraduate major areas. In the study, the weights were used both to assess relative contribution to separation and to give meaningful interpretation to the resulting LDFs. Two-dimensional plots of group centroids were used in both of these studies as an aid to interpretation. A third study which used discriminant "weights" was reported by McNeil (1968). What weights these were was not made explicit. Here, 521 sixth grade children in four subcultural groups were considered for separation by six factors which resulted from a "factor analysis" of 20 semantic differential scales.

Discriminator-LDF correlations were utilized in two studies for purposes of LDF interpretations. Field, et al. (1971) obtained measures on 57 undergraduate Australian students using an 18-item questionnaire assessing teaching behavior. These data were examined to evaluate discrimination among six teachers, including one "ideal" teacher. Substantive interpretations of the LDFs were made. Bausell and Magoon (1972) used data on 29 items of the Purdue Rating Scale for Instruction for approximately 2000 sophomore to senior university students for purposes of differentiating four criterion groups defined by student expected grades. Total-group variable-LDF correlations were used to substantively interpret the LDFs, as well as to assess the "importance" of the discriminators. They also incorporated the statistic, $1 - \Lambda$, for interpretive purposes. See also, Whellams (1973).

The discriminant analysis techniques used by Chapin (1970) were not clear. In his study of four groups of mathematics teachers (determined by

principal ratings) on whom personal and academic measures were obtained, he states, "After extensive sorting through 40 variables, it was found that only a few variables contributed significantly to the discriminant analysis [p. 161]."

The two remaining studies to be briefly reviewed use a large number of criterion groups. Baggaley, Isard, and Sherwood (1970) used 17 groups (14 academic, three "miscellaneous") of university juniors; ten personality measures were obtained on each of 628 students. These investigators examined "normalized" vector coefficients for relative variable contribution, and to meaningfully interpret the LDFs. A two-dimensional (why two?) plot was given. A set of 26 personal and academic measures for college undergraduates was used by Burnham and Hewitt (1972) to differentiate among 16 occupational groups in a follow-up study. Univariate Student t-values were considered to assess relative contribution of the discriminators.

From a statistical point of view, criticisms of the methodology used or of the reporting in some of these studies are possible. It is recognized, however, that writers and/or editors may have reasons for not including all the details of the techniques used.

Estimation

To restate, estimation is that aspect of discriminant analysis that pertains to characterizing inter-group distance and strength of relationship.

Measures of Distance

About 1920 K. Pearson proposed his coefficient of racial likeness (CRL) as a measure of distance which was subsequently used mostly in craniology. In the middle to late 1920's G. M. Morant suggested a corrective factor to be applied to the CRL to offset effects due to varying sample sizes; at about the same time, P. C. Mahalanobis proposed a Euclidean distance measure. [See Hodges (1950, pp. 5-25) for a more complete development of the history of distance measures.] The distance between two population centroids may be expressed as

$$\Delta = [(\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)]^{\frac{1}{2}} ,$$

where μ_k is the centroid of population k, and Σ is the covariance matrix common to the two populations. The quantity Δ has become known as Mahalanobis' generalized distance, and the square of the sample distance,

$$[9] \qquad D^2 = (\underline{\bar{X}}_1 - \underline{\bar{X}}_2)' S^{-1} (\underline{\bar{X}}_1 - \underline{\bar{X}}_2)$$

is often referred to as Mahalanobis' D^2 statistic. The (pxp) matrix S in [9] is defined by $(N_1 + N_2 - 2)\, S = E$; $\underline{\bar{X}}_k$ is the centroid of group k. Although S is an unbiased estimator of Σ, it should be noted that D^2 is not an unbiased estimator of Δ^2 (Rao, 1949). Since an unbiased estimator of Δ^2 often results in negative estimates of the square, its use is discouraged. In presenting a logical derivation of D as a distance

measure, Rao (1952) points out that D "...is applicable only to groups in which the measurements are normally distributed [p. 355]."

If significant group separation is found, it is possible to gain some insight regarding group differences by simply calculating the Euclidean distance (as in [9]) between all pairs of centroids. If, for example, distances between all pairs of K-1 of the groups are small, yet at the same time, the kth group is distinctly separated from the other K-1 groups, it is clear that the only separation taking place occurs between the kth group and its complement, i.e., the other K-1 groups. These pairwise group distances are often given in the output of computer programs -- e.g., the BMD 7M program (Dixon, 1973).

It may be noted in passing that a transformation of each D^2 statistic may be used as a test statistic in the two-group case. This transformation -- see Rulon and Brooks (1968, p. 69) -- may be considered as an alternative to Hotelling's T^2 or Wilks' Λ statistics.

As will be apparent later, the distance function D^2, or a variation thereof, appears in most multivariate classification schemes. That is, a measure of distance between an individual's data point, $\underline{X}$, and the kth group centroid is of interest. This measure, assuming a common population covariance matrix, Σ, is given by

$$\Delta_k = [\underline{X} - \underline{\mu}_k)' \, \Sigma^{-1} (\underline{X} - \underline{\mu}_k)]^{\frac{1}{2}}$$

and the sample distance is given by

[10] $$D_k = [(\underline{X} - \underline{\bar{X}}_k)' \, S^{-1} \, (\underline{X} - \underline{\bar{X}}_k)]^{\frac{1}{2}} \quad .$$

Rao (1952, p. 257) gives a generalization of Mahalobis' D^2 statistic (labeled "V" by Rao):

$$W = \sum_{k=1}^{K} N_k \, (\bar{X}_k - \bar{X})' \, S^{-1} \, (\bar{X}_k - \bar{X}) \quad ,$$

where $\bar{X}$ is the (px1) vector of predictor means across all K groups. It can be shown that W can be used as a chi-square statistic with p(K-1) degrees of freedom to test the hypothesis of equality of the K population mean vectors. As Rao points out, the W statistic may be partioned into independent chi-square statistics so as to judge the significance of information lost when some variables are deleted. This criterion is equivalent to Hotellings' trace statistic, the use of which was proposed by Miller (1962) for variable selection -- see Friedman and Rubin (1967, p. 1162). The value of W is part of the output of the BMD 5M program (Dixon, 1973). It turns out that when K = 2, $W = D^2 \cdot N_1 N_2/(N_1 + N_2)$.

Measures of Discriminatory Power

As most researchers who have toyed with univariate measures of association know, the numerical values of most proposed indices for a given set of data are nearly the same. This has also been shown to be the case in the multivariate situation by Stevens (1972a) and Huberty (1972a). Multivariate measures of strength of relationship, or of discriminatory power, that have been proposed are : (1) $1 - \Lambda$ [Wilks' statistic], (2) $U/(1+U)$, where $U = \text{tr}\,(E^{-1}H)$, (3) $U'/(1+U')$, where $U' = (N-p-1)U/N$, and (4) an extension of Hays' (1973) omega squared. Tatsuoka (1973b) has studied the properties of this last measure which he proposed earlier (Tatsuoka, 1970) and which may be expressed as

$$\hat{\omega}^2_{mult} = 1 - \frac{N}{(N-K)\Lambda^{-1} + 1} \quad .$$

The intent of Tatsuoka's more recent study was to develop an unbiased estimate of

[11] $$\omega^2_{mult} = 1 - \frac{|\Sigma|}{|\Sigma + A\,A'/K|} \quad ,$$

where Σ is the (pxp) covariance matrix common to the K populations, and A is a (pxk) matrix of effect parameters -- the (j,k)th element of A is the deviation of the kth population mean from the general mean for the jth variable. (Formula [11] expresses the proportion of generalized variance of the p variables attributable to differences among centroids.) Since various attempts to develop such an estimate were of no avail, an attempt was made to develop a formula for correcting the (positive) bias in $\hat{\omega}^2_{mult}$. Empirical results led to a "rule-of-thumb" correction to be used with small samples:

[12] $$\hat{\omega}^2_{corr} = \hat{\omega}^2_{mult} - \frac{p^2 + (K-1)^2}{3N}\,(1 - \hat{\omega}^2_{mult}) \quad .$$

It was found that $\hat{\omega}^2_{mult}$ itself can be used when $\omega^2_{mult} \leq .30$ and $N/p > 100$ or when $\omega^2_{mult} \geq .50$ and $N/p > 50$; however, neither situation is typical of that found in educational research. Formula [12] was deemed to be adequate, at least when $p(K-1) \leq 49$ and $75 \leq N \leq 2000$.

Another index of discriminatory power which has received some attention is the proportion of correct classifications across all K groups (Cooley and Lohnes, 1971, p. 329). In a multiple discriminant situation this indicator of strength of predictive validity may be more appropriate than a correlative measure. This statistic will be discussed in some detail later.

Classification

As noted previously, the original intended purpose of "discriminant analysis" and the linear discriminant function (LDF) was that of classification. Given a sample of individuals (or objects) from each of two or more populations, we want to construct a method of assigning a new individual to the correct population of origin on the basis of measures on p variables. Classification procedures are used to solve prediction problems; given measures on the p predictors it is of interest to predict membership in one of the natural or preexisting groups. A more formal view is that classification is used to answer the question: Given an individual with certain measurements, from which population did he emanate? In this sense, the problem of classification may be considered a problem of "statistical decision functions" (Anderson, 1958, p. 126). The p predictors are the independent variables and the single criterion is the grouping variable, the latter being, of course, nonmetric. Illustrations of the use of multivariate classification in behavioral research are given later.

Requisite Information

Various methods of multivariate classification have been proposed. The use of these methods presupposes that the user has knowledge of certain information. This information may in the form of: (1) the density function which best describes the data on hand, (2) restrictions on data conditions necessary to select the most appropriate method, (3) prior probabilities of group membership, and (4) misclassification costs.

The early work in discriminant analysis, specifically that on LDFs and generalized distance, was based on multivariate normal probability distributions. The general theory of classification is not, however, dependent

upon multivariate normality. General distribution-based rules of assigning individuals to populations (so that probabilities of misclassification are minimized) are discussed by Anderson (1958, pp. 142-147) and Overall and Klett (1972, Ch. 12). Parametric and nonparametric density estimators were used to estimate the non-error rate of an arbitrary classification rule and to construct a rule which maximized estimated probability of correct classification by Glick (1972). Insofar as could be determined, little work has done with classification rules involving continuous predictors other than those based on normality.

Assuming multivariate normality, a linear classification rule may be used when it can be further assumed that the condition of equal covariance structures across the K groups is met. As will be shown later, differences in covariances, as well as differences in means, can be utilized in making predictions about group membership, and in estimating error rates.

A helpful consideration to be taken when confronted with the problem of classification is that of prior probabilities of group membership. Such a probability is that of drawing at random an individual of each group from the total population of all K groups. Taking the approach of frequentists, these priors are relative frequencies of individuals of each of the K populations in the total population. From sample data, then, these probabilities are estimated by N_k/N, $k = 1,\ldots,K$. [Tatsuoka (1971, pp. 225-226) discusses problems in using such estimates.] Priors may also be estimated by using Markov models as suggested by Lohnes and Gribbons (1970); alternatives to these suggestions were considered in an empirical study by Lissitz and Henschke-Mason (1972).

Typically, in educational research differential costs of misclassifying individuals into the K groups are ignored -- ignored in the sense

that equal costs are assumed. This is not too surprising since quantification of costs of misclassification in educational research may be difficult, even though only relative costs are important.

Classification Rules

Various parametric and nonparametric rules of classification have been proposed. In most of these rules some notion of "distance" comes into play; that is, an individual is assigned to that group whose centroid is closest to the data-point representing him. "Closeness" is measured by a probabilistic notion of "distance," as opposed to the geometric Euclidean distance measure discussed in an earlier section. The use of the LDF for classification purposes in the two-group situation was initially based on simple Euclidean distance -- assuming multivariate normality and equal covariance structure, an individual was assigned to the group with the mean discriminant score nearer to his discriminant score. Fisher's LDF as a classification statistic was not at first considered in reference to a probabilistic model. The relationship of posterior probability of group membership to the LDF was noted by Welch (1939) when he proved that the assignment procedure based on the LDF minimizes the probability of misclassification under certain restrictions. Von Mises (1945) extended Welch's notions to the K-group case, and removed the restriction that probabilities of misclassification per group be equal.

The classification statistics discussed in this paper will be stated in terms of estimates of population parameters; in so stating, no claim is made that an optimum solution is obtained. Further, only the situation of equal costs of misclassification will be considered. Assuming multivariate normality and identical population covariance matrices, the

distance measure [10] has been used as a classification statistic -- as well as a criterion in cluster analysis (Friedman and Rubin, 1967). An individual, with score vector $\underline{X}_i$ is assigned to that group, k, for which the distance measure

$$[13] \qquad D_{ik} = [(\underline{X}_i - \underline{\bar{X}}_k)' \, S^{-1} \, (\underline{X}_i - \underline{\bar{X}}_k)]^{\frac{1}{2}}$$

is least. These measures may be transformed to "centour scores" which are functions of probabilistic distances (Cooley and Lohnes, 1971, p. 265). For a given individual, the assignment is based on the largest centour.

An inadequacy of [13] is that differential prior probabilities, p_k, of group membership are ignored. Using the multivariate normal distribution function and retaining the equal covariance condition a modification of [13] becomes

$$[14] \qquad L_{ik} = -\tfrac{1}{2} \ln|S| - \tfrac{1}{2} D^2_{ik} + \ln p_k \, .$$

The more popular form of a "linear discriminant score" (Rao, 1965, p. 488),

$$[14a] \qquad L_{ik} = \underline{\bar{X}}'_k \, S^{-1} \underline{X}_i - \tfrac{1}{2} \underline{\bar{X}}'_k \, S^{-1} \underline{\bar{X}}_k + \ln p_k \, ,$$

is equivalent to [14], since the terms $-\frac{1}{2}\ln|S|$ and $-\frac{1}{2} \underline{X}'_i \, S^{-1} \underline{X}_i$ are common to all k. Thus, individual i is assigned to that population whose corresponding sample yields the largest value of the classification statistic [14]. Such a rule minimizes the number of misclassifications, in a parameter sense, and is equivalent to a rule which assigns the individual with measures $\underline{X}_i$ to that population for which the posterior probability of population membership is largest. Some writers [e.g., Eisenbeis and Avery (1972, p. 18)] prefer to express the classification statistic as a posterior

probability:

$$[15] \qquad P_{ik} = \frac{P_k \exp(-\frac{1}{2} D_{ik}^2)}{\sum_{k'=1}^{K} P_{k'} \exp(-\frac{1}{2} D_{ik'}^2)} .$$

[Statistics [14] and [15], which are equivalent in the sense of classification results, are, in turn, equivalent to those reported in Rule IV by Cooley and Lohnes (1971, p. 269) and in Rule 5.6 by Eisenbeis and Avery (1972, p. 19).] Expressions [14] and [15] lead to what is sometimes referred to as the "linear classification rule"; [14] is linear in that L_{ik} is linear in $\underline{X}_i$. Equation [15] exemplies the Bayesian conditional-probability model.

Another linear rule based on posterior probabilities of group membership under the present conditions has been proposed. The formula used to compute the posterior probabilities is based on "Case E. $\Sigma_k = \Sigma$ but unknown, $\underline{\mu}_k$ unknown," presented by Geisser (1966, p. 155). [See also Cooley & Lohnes (1971, p. 269).] Geisser's work resulted in the classification statistic,

$$[16] \qquad Q_{ik} = \frac{P_k \cdot h_{ik}}{\sum_{k'=1}^{K} P_{k'} \cdot h_{ik'}} ,$$

where h_{ik} is the "predictive density of a future observation (vector) given the available data," and is proportional to

$$\left[\frac{N_k}{N_k + 1}\right]^{p/2} \left[1 + \frac{N_k D_{ik}^2}{(N_k + 1)(N - K)}\right]^{-(N-K+1)/2} .$$

It has been shown (Huberty, 1971b) that when the N_k-values are identical statistics [14] and [16] yield the same results. Of course, since for a given individual the denominators of [15] and [16] are constant, only the maximum values of the numerators need be considered in making assignments. However, actually obtaining the probabilities may provide information in addition to that of mere number of correct and incorrect classifications. For example, a vector of P_{ik}- or Q_{ik}- values of (.80, .15, .05) versus a vector of (.48, .46, .06) would lead to the same decision, namely, assign to group 1. However, it may be informative to examine such vectors to determine those individuals, and their characteristics (as reflected in $\underline{X}$-vectors), who are misclassified. Also, by examining the probability vectors, it can be determined the group that an individual is most like (highest value) and the group he is most unlike (lowest value).

Under the condition of unequal covariance matrices, variations of the above three classification statistics are called for. If equal covariance structure cannot be assumed, then S in [13] is replaced by the sample covariance matrix for each group:

$$[17] \qquad D'_{ik} = [(\underline{X}_i - \underline{\bar{X}}_k)' \, S_k^{-1} \, (\underline{X}_i - \underline{\bar{X}}_k)]^{\frac{1}{2}} \quad .$$

Taking into account different group covariance matrices, the counterpart of [14] may be expressed as a "quadratic discriminant score,"

$$[18] \qquad L'_{ik} = -\tfrac{1}{2}\ln|S_k| - \tfrac{1}{2}(D'_{ik})^2 + \ln p_k \quad .$$

Again, this classification statistic may be transformed to a statistic that yields posterior probabilities of group membership:

[19] $$P'_{ik} = \frac{p_k \, |S_k|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(D'_{ik})^2\right]}{\sum_{k'=1}^{K} p_{k'} |S_{k'}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(D'_{ik'})^2\right]}$$

[See Tatsuoka (1971, p. 228) for a discussion of the transformation of [18] to a posterior probability.] Formulas [18] and [19] lead to what is sometimes called the "quadratic classification rule," since [18] is quadratic in $\underline{X}_i$. A second quadratic rule -- [18] and [19] yield identical results, as do [14] and [15] -- has been proposed that is also built on posterior probabilities of group membership. The probabilities are based on a Bayesian density specified in Geisser's (1966, p. 154) "Case C. Σ_k unknown, $\underline{\mu}_i$ unknown." (See also, Press, 1972, p. 375.) The posterior probabilities are given by

[20] $$Q'_{ik} = \frac{p_k \cdot g_{ik}}{\sum_{k'=1}^{K} p_{k'} \cdot g_{ik'}} ,$$

where g_{ik} is a density proportional to

$$\left[\frac{N_k}{N_k + 1}\right]^{p/2} \Gamma\left(\frac{N_k}{2}\right) \left[1 + \frac{N_k (D'_{ik})^2}{N_k^2 - 1}\right]^{-N_k/2} \Bigg/ \Gamma\left(\frac{N_k - p}{2}\right) \left|(N_k - 1)S_k\right|^{\frac{1}{2}} .$$

It can be shown that [20] and [19] (and hence [18]) yield identical results when the N_k -values are the same.

Horst (1956a) considered a formulation of the classification problem involving separate regression equations contrasting each criterion group in turn with all others. In finding the regression equation corresponding to group k, the dichotomous criterion variable assumes the value 1 for individuals in group k and 0 otherwise. To estimate the coefficients used in Horst's "least squares" multiple classification method, the total covariance matrix of the predictors is involved. The following classification

statistic results, when N_k-values are identical:

$$[21] \qquad Y_{ik} = \underline{b}_k' (\underline{X}_i - \overline{\underline{X}}) + \overline{Y}_k \quad ,$$

where $\underline{b}_i$ is the (pxl) vector of sample coefficients for group k,

$$\underline{b}_k = T^{-1} \underline{u}_k \quad ,$$

with T = H+E and $\underline{u}_k$ the vector of deviation score cross-products of the predictors and the (dichotomous) criterion, the deviations being taken from the grand means, and $\overline{Y}_k = N_k/N = 1/k$. The statistic [21] leads to a decision rule which assigns an individual to that population for which his corresponding composite score is nearest unity. A modification of [21] is required with different N_k-values (Horst, 1956b).

For ease of reference the seven statistics presented are given in Table 1.

--Insert Table 1 About Here--

With this apparent variety of classification statistics available, which does one use? Assuming the condition of multivariate normality is tenable, the choice seems to depend upon whether or not the added condition of equal covariance structure is also tenable, and whether or not differential priors are to be involved. [An added criterion of choice may be one's preference for use of statistics based on the classical approach or on the Bayesian solution of Geisser (1964, 1966) and Dunsmore (1966). The Bayesian solution is simpler to come by in that it is not based on any complicated distribution theory.] In a Monte Carlo study where both noncross-validation and cross-validation results were reported Huberty and Blommers (1974) concluded that the rule based on [21], or its modification for unequal N_k-values does not yield as great accuracy as that

yielded by the other rules considered. Knutsen (1955), however, concluded from a single sample that this rule was more accurate than the rule based on [14]. Huberty and Blommers (1974) also found that by incorporating prior probabilities into a rule, classification accuracy is enhanced -- rules based on [13] or [17] versus those based on [14] or [18]. They further concluded that rules based on [16] and [19] yielded nearly the same results; no comparison of [16] and [15] was made, but since sampling was made from populations with a common covariance matrix, it is conjectured that these two statistics would yield similar accuracy of classification. So, in the linear case -- when covariance matrices are taken to be equal -- either [14] (or [15]) or [16] may be used as classification statistics with expected results very similar.

Insofar as could be determined no studies have been undertaken to compare the efficiency of [18] (or [19]) to that of [20]. Cooley and Lohnes (1971, pp. 270-272) report the results of some Monte Carlo classification studies in which the efficiencies of [14], [16], and [19] are compared. (Their "Anderson method" is equivalent to that based on [14].) Their results reported do not suggest the superiority of any one of the three statistics; they do conjecture, however, that the rule based on [19] "...might suffer more from capitalization on chance differences in covariances [p. 272]." It is noted in passing that the equivalence of [14] and [16] with equal N_k-values and $p_k = N_k/N = 1/K$ for all k (Huberty, 1971b), was empirically verified by Cooley and Lohnes (1971, p. 272) when these statistics led to the same proportion of correct classifications.

A description and application of a simulation program designed to obtain estimates of different types of misclassification probabilities and to compare linear and quadratic classification rules is given by

Michaelis (1973). The different misclassification probabilities are those discussed in a subsequent section, "Estimating Error Rates." It was assumed that all prior probabilities and misclassification costs are equal. The two multivariate normal classification rules used were, in essence, those based on [14] (the linear rule) and on [18] (the quadratic rule), disregarding the $\ln p_k$ terms. In the simulation process, the model parameters were chosen to be equal to parameters which had been estimated from real data. The basic model considered was one where $K = 5$ and $p = 8$ with unequal population covariance matrices. Sample sizes of 30 and 100 per group were used. Both "internal" classification," where the parameter estimates are based on the samples classified, and "external classification," where the parameter estimates are based on a sample other than that classified, were used. As might be expected, quadratic classification yielded the better results. The difference between the results of internal and external classification was found to be substantially larger for the quadratic than for the linear rule, especially for the smaller sample size. This is presumably due to the fact that the number of estimated parameters is much smaller in the linear rule. For all simulated larger samples ($N_k = 100$) the external quadratic classification gave better results than the corresponding linear classification, although the estimation of the parameters was not yet very good, as could be seen from the differences between internal and external results -- especially for quadratic classification. Even with the smaller sample sizes, where the differences between internal and external analysis are very large, in most samples external quadratic classification gave better results than the corresponding linear classification. In conclusion, Michaelis recommends both an internal and an external classification in each practical application. The differences between the two resulting proportions indicate an interval in which the "true error" can be expected

to lie. Further, if the two proportions differ greatly, one could expect to achieve better classification of independent samples by increasing the sample size. Results of several other simulation experiments are graphically reported. This is an excellent reference for anyone interested in simulation experiments in multivariate classification.

A few writers have advanced arguments in favor of using classification statistics based on LDFs rather than on the original predictors (Cooley and Lohnes, 1962, p. 139; Tatsuoka, 1971, p. 232; Eisenbeis and Avery, 1972, p. 56). Briefly, the arguments presented for using such "reduced space" procedures are: (1) the linear transformation (when covariance matrices are equal) preserves the overall structure as well as distances in the reduced (or discriminant) space of dimension $r = \min(K-1,p)$, and computations are easier; (2) since most often, $r \leq 2$, computations are further reduced, and interpretations are simpler; (3) the Central Limit Theorem implies that the distribution of the linear discriminant scores for each group approaches normality; and (4) classifications may be more consistent over repeated sampling because of relatively greater stability of statistics based on LDFs.

In their empirical study, Huberty and Blommers (1974) found that the decision rule based on [19] with discriminant scores as input did better over repeated sampling than with original predictor scores. From the results of another empirical study, where the conditions of normality and equal covariance matrices were controlled, Lachenbruch (1973) concluded that a reduced space classification method works about as well as the method based on [14] if the population means are collinear or nearly so. Otherwise, [14] proved much better. In that study, the sample size and the p_k-values were taken to be equal across the groups. Lohnes (1961) used [17] in

classifying three sets of real data in both the original spaces of the predictors and in the discriminant space. The equal covariance structure condition was not met for at least two sets -- results were not presented for the third data set. For all three sets it was concluded that the two methods produce comparable classification results.

Four different classification rules were investigated in an empirical study -- using data on engineering students -- by Molnar and Delauretis (1973). The statistics used may be expressed as (1) [17] in the discriminant space, (2) [19] in the discriminant space, (3) [14] with equal p_k-values [the equivalent of the statistic used in the BMD 5M program (Dixon, 1973)], and (4) [15] which is equivalent to that used in the BMD 7M program. The second statistic yielded slightly better results than the first for one set of data. The first, third, and fourth statistics did equally well for a different data set. The purpose of such comparisons is not clear; conclusions about the relative efficiencies of the rules cannot be made from such a study. For the first set of data involving three groups, three two-group classification analyses were also carried out.

In addition to discussing the use of LDFs in classification, Overall and Klett (1972, Ch. 14) indicate that another orthogonal transformation may be useful for classification purposes. The transformation is obtained via a principal components analysis of the within-groups covariance matrix, S. The use of two LDFs and four principal components were compared for a set of data involving three criterion groups, 16 predictors, and nearly 3000 individuals. Results of maximum likelihood classification -- as from [15] except that priors are not considered -- applied in the two reduced spaces (of two and four dimensions) were very similar.

Much more empirical work needs to be done in the multi-group case of

assessing the efficiency of the various classification statistics under different conditions. Comparisons within the set of linear rules, within the set of quadratic rules, and across the two sets remain problems for future study, as do those involving the use of different priors and non-normal distributions. Eisenbeis and Avery (1972, p. 53) conjecture that the use of linear versus quadratic techniques will affect the classification less than variation in prior probabilities. In a two-group study Anderson and Bahadur (1962) pointed out that deviations from normality may affect the results of quadratic classification much more than those of linear classification. The study of reduced space classification using different orthogonal transformations of the raw data in the dimension reduction may also be of interest.

Efficiency of Classification

The results of any classification analysis may be summarized in a KxK classification table (or "confusion matrix"). The two dimensions of the square matrix are actual group membership and predicted group membership. One set of diagonal elements of such a cross-tabulation matrix give the number of "hits" or correct classifications for each group. Data from this matrix may be used to test whether the classification procedure used is significantly better than a purely random partitioning of the decision space; i.e., better than if assignments of individuals to groups were based on chance alone. Since the only entries in the confusion matrix of interest for this test are those on one of the diagonals, the usual Pearson chi-square test is not appropriate. Significance by this test is a necessary but not sufficient condition for concluding that the number of correct classifications is greater than would be expected by chance.

Three statistics have been proposed for testing the efficiency of

a classification procedure; the referent distribution that may be used for all three is the standard normal. They are reported by Lubin (1950), McHugh and Apostolakos (1959), and Press (1972, p. 382). (The second reference has a minor error in a formula used.) None of these tests is strictly appropriate since the same data are being used to test the procedure as to define the procedure. If sufficient data are available, it might seem more appropriate to use a holdout sample to assess efficiency; however, as will be noted in the next two sections, better methods are available. The hit rate yielded by these better methods may then be compared to the expected hit rate based on chance alone, $\Sigma\ p_k\ (N_k/N)$.

The efficiencies of two different classification procedures applied to the same data may be compared via McNemar's test of related proportions. This test, and an extension of it proposed by W. G. Cochran for use in comparing more than two procedures, are discussed by Hays (1973, pp. 741, 773).

Estimating Error Rates

Most of the work done with methods of estimating proportions of classification errors deals with the two-group situation. Much of this research will be reviewed in the next section.

Three types of errors may be associated with a classification rule: (1) true error, (2) actual error, and (3) apparent error (see Hills, 1966). True (or optimal) error is the long-run frequency of misclassifications using a classification rule which assumes population parameters are known. Actual error is the long-run frequency of misclassification using a rule which uses estimates of the unknown parameters. Apparent error is the

proportion of the "norming sample" misclassified by a rule which uses parameter estimates -- internal classification. As will be shown in the next section, estimators of true error and of actual error are simply formulated in the two-group case; however, the formulation for estimators in the multi-group case are complicated, indeed (Glick, 1972). For large samples, true error and actual error will be approximately equal, and an estimator of one could be used for the other. Apparent error has often been used as an estimate of these two types of error. As might be intuited, since classifying the norming sample with a rule determined by this same sample is quite likely to capitalize on chance, apparent error may grossly underestimate actual or true error.

A better estimate may be obtained by extending a technique (Lachenbruch, 1967) which was proposed for the two-group case. This ("jackknife") technique requires the application of a classification rule N ($=\Sigma N_k$) times, withholding a different vector of measures each time. The individual whose vector was withheld is then reclassified using the statistics based on the other N-1 sets of measures. The proportions of misclassified individuals from each group are used as estimates of the conditional probabilities of misclassification. One minus the proportion of misclassification across all K groups may be used as a measure of the discrimnatory power of the predictors. Such a measure informs a researcher how well a set of predictors differentiates the criterion populations, and it may serve as a yardstick in determining whether the addition of new variables or the deletion of old ones is warranted (Geisser, 1970, p. 60).

Classification in Two-Group Case

Rather than considering two linear discriminant scores (i.e., values of L_{ik} in [14a]), only one comparison is involved when there are only

two criterion groups. The classification decision can be made by computing

[22] $$L_{i1} - L_{i2} = (\bar{X}_1 - \bar{X}_2')\, S^{-1}\, X_i - c,$$

where

$$c = \tfrac{1}{2}(\bar{X}_1'\, S^{-1} \bar{X}_1 - \bar{X}_2'\, S^{-1} \bar{X}_2) + \ln p_2 - \ln p_1 .$$

It is noted that the vector of coefficients of the single LDF (see [3]) may be taken as

$$v_1' = (\bar{X}_1' - \bar{X}_2')\, S^{-1} .$$

The decision rule is to assign individual i to the first population if $v_1' X_i \geq c$, and to the second population if $v_1' X_i < c$. The formal equivalence of two-group discriminant analysis and multiple regression analysis, where the criterion variable is measured by group membership, was mentioned previously in this paper. When the number of individuals in each of the two groups is the same, classification based on [22] is identical to that based on [21] for K = 2.

It is in the two-group case where most work has been done in assessing the robustness of linear classification rules to various departures from assumptions. Investigations of linear rules for unequal covariance matrices have been performed by Kossack (1945), Smith (1947), and Gilbert (1969). Studies involving the classification of non-normal data are reviewed in a later section. See also, Lachenbruch (1966).

The use of prior probabilities or "base rates" in a univariate classification scheme was used about twenty years ago by Meehl and Rosen (1955) in a two-group study. This use of unequal priors to increase classification accuracy was critiqued by Cureton (1957). Overall and Klett (1972, pp. 265-267) discuss unequal base rates used jointly with LDF scores in

graphically determining appropriate cut-off points for classification into one of two groups. Alf and Dorfman (1967) determine a cut-off score for a single predictor or a weighted sum of predictors such that the expected value of the decision procedure is maximized, taking gains and losses associated with correct and incorrect assignments into account. See also, Gregson (1964).

Considerable theoretical and empirical research has been reported that deals with estimating probabilities of misclassification in the two-group case. Hills (1966) has given an excellent account of the problems involved in estimating various error rates in multivariate two-group classification problems. Hockersmith (1969) reviews numerous methods of estimating true and actual error -- refer to the preceding section -- and reports the results of a Monte Carlo study comparing the accuracy of the methods. The comparative accuracy of the methods depends upon the number of predictors, group size, and distance between the two population centroids. It was generally concluded that a method using a holdout sample, where a subset of the original set of observations is classified using the rule determined by the remaining observations (the norming sample), was inferior to the others. As might be expected, it was concluded that apparent error was a poor method in nearly all situations studied. If one method were to be selected when the normality condition is questionable, it would be the "U method" suggested by Lachenbruch (1967) which was described in the preceding section of this paper. If the normality condition is met, a method which combines the features of the U method and the use of the normal distribution is recommended; here an estimate is taken as

[23] $$\tfrac{1}{2}\left[\Phi\left(-\bar{L}_1/s_{L_1}\right) + \Phi\left(\bar{L}_2/s_{L_2}\right)\right] ,$$

where Φ is the standard normal distribution function, $\bar{L}_k$ is the mean of the N_k values of [22] in group k each based on $N_1 + N_2 - 1$ observations, and s_{L_k} is the standard deviation of such values for group k. Other specific conclusions were reached which, along with the general conclusions, were comparable to those of Lachenbruch and Mickey (1968) in a somewhat similar study. See also the three papers by Sorum (1971, 1972a, 1972b).

Results of studies by Lachenbruch (1968) and Hockersmith (1969) have also led to conclusions regarding sample size. The recommendations made are dependent upon the number of predictors, the distance between the two populations, and the tolerance between the estimated and optimum error rate. Tables are provided by both writers which indicate a desired common sample size in different situations. Using [23] as an error rate estimate, sample size requirements are summarized by Lachenbruch (1968) as follows: (1) for large tolerance only small samples are needed; small tolerances imply the need for large samples; (2) groups widely separated need smaller samples for classification than groups that are close together; and (3) as the number of parameters increases, the required sample size to number of parameters decreases. Hockersmith (1969) draws similar conclusions, and specifically states that for the better error rate estimates, "...a sample size of 40 in each group could be used to insure with some confidence that the estimate of (true) error will be within a tolerance of .05 [p. 80]." See also, Dunn (1971).

Specific Uses of Classification

Classification has at times proved helpful when used in conjunction with other multivariate data analysis techniques. A use of classification procedures in a prediction study is given by Lissitz and Schoenfeldt (1974);

probabilities determined by [19], with equal priors, were used as weights in a multivariate prediction model. Rogers and Linden (1973) transformed values obtained from [13] to probabilities of group membership that were used as classification statistics so as to test the efficiencies of three grouping (or clustering) methods on a given set of data. A classification procedure was also used by Schoenfeldt (1970) to validate a clustering method. It is noted that the purpose of this latter type of study is to define groups rather than to predict group membership; thus the usual tests of significance used in discriminant analysis do not apply (Friedman and Rubin, 1967, p. 1167).

Jackson (1968) studied two methods of estimating unknown values in discriminant analysis and used as his criterion of comparison the proportion of correct classifications (one minus apparent error) yielded by each method. The criterion of one minus apparent error was also used by Huberty (1971a) in assessing the effectiveness of various methods of selecting a subset of predictors of a given size.

The close relationship of multivariate classification techniques to "profile analysis" is pointed out by Overall and Klett (1972, Ch. 15).

Classification Research Applications

As in the section on discrimination applications, only selected journal articles in behavioral research dealing with applications of classification procedures will be reviewed. Two sets of studies are reviewed: (1) studies dealing almost exclusively with classification, and (2) studies using both discrimination and classification techniques.

Only one study selected (Doerr and Ferguson, 1968) carried out classification in the reduced space. Eight vocational test scores and five interest inventory scores for 9[illegible] high school students were used for assignment of

individuals into one of eight vocational course groups. The null hypothesis of MANOVA was rejected. Two "significant" LDFs were determined by examining the ratios of the individual eigenvalues of $E^{-1}H$ to their sum, i.e., to the trace of $E^{-1}H$. A random 10% (from each group) was used for cross-validation purposes; both internal and external classification results were reported. The classification statistic used was not indicated; presumably it was [19] in the reduced space -- no priors were specified.

In some studies, conducted for the purpose of prediction where the dependent measure is nominal, the classification statistic is not made explicit. In a study by Stahmann (1969) it was merely stated that, "Multiple discriminant analysis was used as a classification procedure ...[p. 110]." That study involved approximately 500 bachelor degree graduates in five fields of study; ten academic test scores, nine occupational interest inventory scores, and two self expressions of major field were used as predictor measures. Neither equality of covariance matrices nor centroids was considered. A holdout sample was employed for validation purposes; the proportion of the original sample was not specified. External classification results were reported for one data set, while internal classification was used for two other sets of data. Fourteen measures, seven of which were academic test scores, on 160 college freshmen were used by Chastian (1969) to predict membership on one of four classes, two audio-lingual and two "cognitive." The hypothesis of equal mean vectors was rejected. Total and separate group correlation matrices were given. Internal classification results via an unknown statistic were tabled; the need for external classification was noted, however. Pearson's chi-square statistic was used to assess the efficiency of classification. Multiple regression techniques were used with the same predictors, but to answer a different question.

Four "intellective" and 30 "nonintellective" variables were used by

Keenen and Holmes (1970) in predicting membership in one of three groups (graduates, withdrawals, failures) of 364 college freshmen. The classification statistic used is [17]. A 50 percent holdout sample was used to validate the classification procedure; internal and external classification results were reported. The correlational statistic, $1-\Lambda$, was considered since its use "...is presently felt to be more meaningful than F (a transformation of Λ) in evaluating the results of a discriminant analysis [p. 93]."

See Alumbaugh, Davis, and Sweney (1969) and Cohen (1971) for examples of two-group classification studies.

Five studies will now be mentioned that utilized both discrimination and classification techniques; for these studies only the statistical techniques used will be discussed. All but one of the studies reported only internal classification results. The classification statistic was not specified by Kirkendall and Ismail (1970), Southworth and Morningstar (1970), or Asher and Shively (1969). In the first study Wilks' Λ was used to test for centroid differences among the three populations. Standardized weights were used to assess variable contribution to the lone LDF which was retained due to the percent of "total among-group variation" absorbed. The generalized Mahalonobis distance statistic -- labeled W in the present review -- was used in the second study to test the null hypothesis of MANOVA. It was not specified which LDF weights were used to sort out the two most effective discriminators. Asher and Shively substituted mean values of variables within each of their four groups for missing data. Three LDFs were statistically significant, but only two were considered since the two associated eigenvalues accounted for nearly 97% of the trace of $E^{-1}H$. The type of weights used for interpretive purposes was not made explicit.

Standardized weights were used by Neal and King (1969) and Wood (1971). Following the use of Wilks' lambda statistic, Neal and King carried out both an internal and external classification using, presumably, statistic [19] -- priors were not specified. A "chi-square Goodness of Fit test" was used to determine if the observed distributions, via both internal and external classification, could have been obtained by chance. The results of the statistical classification were compared to those of a "configural analysis" by means of a "chi-square contingency test." The BMD 5M program was used by Wood for his group assignment procedure; that is, statistic [14] lacking the $\ln p_k$ term. The coefficients of the classification equations, [14a], were "scaled" to determine the "relative discriminatory power for each variable."

Other Issues, Problems, Developments

Regression Analysis and Classification

The formal relationship between multiple regression analysis and two-group discriminant analysis was noted previously in this paper. There have been a few studies which have attempted to compare the classification efficiencies of the two methods on a given set of data. In one study (Alexakos, 1966), college grade-point average was used as the criterion measure for both analyses; in two other studies (Dunn, 1959; Bledsoe, 1973) the criterion measure is different for the two analyses. In the first study the classification method used was not made clear; Dunn used the statistic given in [13], while Bledsoe used [15]. From a statistical viewpoint, the appropriateness of such comparisons appears questionable. Use of the same criterion measure in both analyses would ignore requirements for one or the other; also, the results of such a comparison could be different depending upon the classification statistic used. Using different criterion

measures implies that the statistical predictions would not be comparable. Some substantive knowledge may be gained from such comparisons, however. Of course, the two analyses answer different and, perhaps, interesting questions (Tiedeman, 1951). [See Rulon, et al. (1967, pp. 323-336.]

A hybrid of the regression and discriminant analyses which has considerable intuitive appeal is a "joint probability model" which was originally proposed by Tatsuoka (1956). This model considers information concerning group membership in combination with that concerning success or productivity in a group. This is an extension of the classification problem in terms of applications to vocational and educational guidance. The approach is discussed and illustrated by Tatsuoka (1971, pp. 237-242) and Rulon, et al. (1967, Ch. 10). As attractive as this approach may appear, it has not enjoyed widespread use; the Lissitz and Schoenfeldt (1974) study referred to earlier used the idea, though it was not very helpful.

Non-Normal Data

Although most research in discriminant analysis using non-normal data has dealt with classification, some work has been done in the area of discrimination. Variable selection was the concern of Elashoff, Elashoff, and Goldman (1967) with dichotomous discriminators in the two-group case, and of Hills (1967) with dichotomous and polychotomous variables in the case of two or more groups. The selection procedure used by Hills was based on one of the nearest neighbor allocation rules of Fix and Hodges (1951); this latter report gives some of the first work in nonparametric discriminant analysis. A theoretical paper by Raiffa (1961) deals with the problem of (sequentially) selecting from items which are scored 0-1, a subset which will discriminate two groups of individuals about as well as the original set.

General non-parametric or distribution-free, as well as specific discrete and other non-normal, univariate and multivariate classification procedures have been very adequately reviewed by Das Gupta (1973). Various procedures have been developed to classify individuals or observations characterized by various types of variables. For example, Solomon (1961) and Cochran and Hopkins (1961) developed classification techniques for categorical variables; Bargmann (1962b) developed a technique to classify time dependent data; Kendall (1966) and Kossack (1967b) developed techniques for ordinal data; Fix and Hodges (1951) developed a non-parametric technique for variables with unspecified distributions. Applications of non-parametric techniques in behavioral research have been very limited; two applications of the analysis of Cochran and Hopkins have been reported by Toms and Brewer (1971) and Kruglick and Brewer (1974). See also, Overall and Klett (1972, Ch. 16).

A multivariate classification procedure which can handle different types of predictor variables has been proposed and illustrated by Henschke, Kossack, and Lissitz (1974). The technique used, which is an extension of that proposed by Kossack (1967a) for the two-group case, accommodates multiple groups and three different types of variables: interval, ordinal, and nominal. It involves the transformation of each variable type in an appropriate fashion so as to convert it to an essentially measureable variable with equal group covariance matrices. The transformation of the nominal variable is based on that used by Bryan (1961). Once the transformations are completed a LDF is formed, and a Bayes classification rule is used. For a single set of data they found their generalized classification procedure to be "clearly superior" to the statistic [16] used on the same data by Lohnes and Gribbons (1970).

Other comparisons of the efficiencies of normal-based and nonparametric classification rules for the K-group case are needed. Empirical studies extending comparisons made in the two-group case by Fix and Hodges (1952), Gilbert (1968), Gessaman and Gessaman (1972), and Moore (1973) would be four possibilities. Another possibility is an extension of the Lachenbruch, Sneeringer, and Revo (1973) study.

Incomplete Data

A number of methods have been proposed for handling the problem of parameter estimation when data values are missing or unknown in a multivariate analysis. Afifi and Elashoff (1966) provide an extensive review of the literature dealing with this problem. The estimation of covariance and correlation matrices for a single population was studied by Timm (1970). Most studies of this kind assume the data are missing at random (see Rubin, 1973).

As in other areas of discriminant analysis, most research dealing with incomplete data has been done for the two-group case. Jackson (1968) presents results of an empirical study where both the number of variables and number of individuals were very large. Her preliminary findings suggest that the far simpler method of using means for missing values gives results comparable with those of an iterative regression estimation technique. Probabilities of correct classification under eight methods of handling (randomly) missing values were studied by Chan and Dunn (1972) using Monte Carlo methods. The mean substitution method (again) and a principal component method were found, in general, to be superior to the other methods for cases considered. These writers caution that their results may not hold up with non-randomly missing data, with non-normal populations,

and with unequal population covariance matrices. For a third study involving only two criterion groups, see Smith and Zeis (1973). See also McDonald (1971).

Use of T Versus E

In an early section of this paper a formulation for obtaining coefficients of LDFs was expressed in terms of the within-groups SSCP matrix, E (see equation [21]). Computationally, it would be equivalent to use

$$|T^{-1}H - \theta I| = 0 \quad ,$$

which would lead to vectors of coefficients which are proportional to those obtained using E (Rozeboom, 1966, p. 562; Porebski, 1966a). In their formulations a few writers prefer the use of T, while most writers use E. The use of the T matrix was suggested by Ottman, Ferguson, and Kaufman (1956) in obtaining classification equations as an alternative to those given by $L_{ik} - \ln p_k$ (see [14a]). The classification statistic proposed by these writers is $\bar{X}_k' T^{-1} X_i - \frac{1}{2}\bar{X}_k' T^{-1} \bar{X}_k$. They claim that one of the principal and unique advantages of such a formulation is that "...once the data for the general population are available, the general population can be further subdivided and more equations developed for an indefinite number of subpopulations [p. 80]." The modified statistic is easily amenable for dropping, adding, or adjusting any of the criterion groups.

In discrimination the important consideration in deciding which SSCP or covariance matrix(ices) to use pertains not to computations but to inferences the researcher wishes to make. Of course, inferential statements are related to the sampling design of the investigation. If the different groups being studied do not represent natural subgroups of some larger populations -- e.g., an experimental study involving different treatments --

then it might seem appropriate to use T. However, when attempting to substantively interpret the LDFs in such a situation the use of T is irrelevant since the LDFs have no population counterpart (Mulaik, 1972, p. 428).

Reporting Discriminant Analysis Results

No matter which purpose or combination of purposes an analysis is to serve, it is recommended that the following be reported: (1) method of sampling, (2) data collection procedures including clear descriptions of measures used, (3) number of individuals in each criterion group, (4) means and variances (or standard deviations) on each variable for each group, and over all groups combined, and (5) the pxp correlation matrix based on E. In addition, the computer program(s) used -- e.g., from a package, or self-written -- should be specified.

When separation is considered, univariate statistics (e.g., ANOVA F-values or the transformed correlational indices such as $\hat{\omega}^2$) should be reported. Some assessment of the equal group covariance structure condition is recommended, such as group covariance matrix determinants or value of a test statistic. The statistic used in testing the null MANOVA hypothesis should be reported -- the type as well as a numerical value.

For discrimination the reporting of the above information plus more is recommended. First of all, the coefficients should be given, indicating whether they are applicable to raw scores or standardized scores. Also, if discriminator versus LDF correlations are used, they ought to be reported, indicating whether they are based on the total-group formulation [5] or the within-groups formulation [6]. If it is inferred that some discriminators could be deleted in subsequent similar studies, coefficients for the retained

variables should be recomputed and reported. If the researcher favors the interpretation of functions beyond that associated with the largest root of $E^{-1}H$, then it is recommended that two-dimensional plots of centroids be presented.

Further, assuming significant separation is determined, reporting an estimate of the proportion of variance of the p variables that is attributable to centroid differences is recommended. Estimates of pairwise group distances (between centroids) using [9] may also be informative.

Certain information ought to be made explicit when reporting results of a classification study. Here too, values used in assessing group covariance structure should be reported. Reporting the classification rule or statistic used is also advised along with the priors used. It is further recommended that a table of hits and misses be given using both an internal and an external classification method.

General References

The sources cited here are restricted to those that can be used as references for discrimination and classification. All of them were referred to at least once earlier in this paper.

To date, the best references, in the opinion of this writer, for discussions on discrimination are not found in books on multivariate methods. Four of these, in order of preference, are Tatsuoka (1973a), Porebski, (1966b), Bargmann (1970), and Bock and Haggard (1968). In the Tatsuoka chapter, issues and problems in interpretation of LDFs are scattered throughout. A very readable discussion of the basic mathematics involved in discrimination and of an approach to interpretation is provided in a pamphlet by Tatsuoka (1970). An elaboration of this coverage is also provided by

the same writer (Tatsuoka, 1971). Brief discussions are given in two chapters by Cooley and Lohnes (1971, Chs. 9 and 12) and at the very end of the fine book by Mulaik (1972). Tatsuoka's interpretations are based on standardized weights, while Cooley and Lohnes and Mulaik prefer the variable-LDF correlations. Eisenbeis and Avery (1972) give a good discussion of the problem of variable selection.

Tatusoka (1971) also provides a coverage of classification procedures, including a good discussion of posterior probabilities of group membership. An excellent general discussion of classifications based on posterior probabilities is given by Overall and Klett (1972), while Eisenbeis and Avery (1972) provide a discussion of the estimation of error rates. For the educational researcher, these latter two books suffer from the drawback of the lack of appropriate illustrations; Eisenbeis and Avery also have some annoying errors in their expressions for a few statistics. The books by Cooley and Lohnes (1971) and Press (1972) are the only ones of those reviewed that present Geisser's classification statistic based on posterior odds -- [16] in the former and [20] in the latter. The book edited by Cacoullos (1973), basically one on multivariate classification, contains at least six very readable papers, all referred to earlier, plus a rather extensive bibliography at the end of the book. The review by Das Gupta is highly recommended.

Computer Programs

One or more of a number of statistical computer "packages" are readily accessible at most institutions -- BMD, OSIRIS, SAS, and SPSS are popular packages. The three "discriminant analysis" programs in the BMD package (Dixon, 1973) have been reviewed quite extensively elsewhere (Huberty, 1974a). The single program in the IBM Scientific Subroutine Package (SSP)

is the same as the BMD 5M program, "Discriminant Analysis for Several Groups." The discriminant "functions" yielded by the BMD multi-group programs are the equations as in [14], not those found via [3].

There are some books that list a number of computer programs (e.g., Veldman, 1967; Cooley and Lohnes, 1971; Overall and Klett, 1972). As can be determined from the description given, the output of Veldman's discriminant analysis program (DSCRIM) includes the variable-LDF correlations as given in [6] and the group centroids. To classify individuals using Veldman's program it is necessary to use his cluster analysis program (HGROUP) which uses the statistic [13]. The discriminant analysis program of Cooley and Lohnes (DISCRIM) yields standardized weights, variable-LDF correlations as given in [5], the value of $1-\Lambda$, and "communalities" as given in [8]. Their classification program (CLASIF) utilizes the method of Geisser [16] for internal classification with prior probabilities defined by group sizes relative to N. The equality of the population covariance matrices is tested with their MANOVA program; however, no quadratic classification results are possible. The discriminant analysis program of Overall and Klett provides output similar to that of Cooley and Lohnes, plus pairwise distance measures (see [9]). The statistic used in their classification program is [15]; internal classification is also possible in a reduced space determined by orthogonal transformations of the original p measures.

A discriminant analysis program is also given by Eisenbeis and Avery (1972). This program, which is reportedly available from The University of Wisconsin for a cost, provides considerably more output information for purposes of variable selection and of classification than those previously mentioned. No variable-LDF correlations are computed, however. The test

of equal covariance matrices is carried out, followed by the use of either a linear [14] or a quadratic [18] classification statistic; reduced space classifications are also optional. The Lachenbruch (1967) jackknife method of estimating the probability of misclassification is used in this program. The combination method of variable selection described in an earlier section is utilized.

A new BMD program, discussed by Dixon and Jenrich (1973), is now available; it requires some special hardware, and may be obtained for a small cost. The program has three very promising added features: provision for (1) more meaningful graphic interpretation of results, (2) the handling of the unequal covariance structure situation, and (3) specifying relative costs of misclassification as well as differential prior probabilities for each group.

There are a few very general multivariate programs (e.g., those by Elliot Cramer and by Jeremy Finn) that are available to users. A program called MUDAID, which is used extensively at The University of Georgia, is an updated version of that by Applebaum and Bargmann (1967); it is now used mostly on the CDC 6400. The Cramer, Finn, and Bargmann programs are used basically for separation and discrimination, with varying outputs. Other individual specific computer programs useful in discriminant analysis are available. References to many programs can be found in the journals, Educational and Psychological Measurement and Behavioral Science.

References

Afifi, A. and Elashoff, R. M. Missing observations in multivariate statistics I. Review of the literature. Journal of the American Statistical Association, 1966, 61, 595-604.

Aitkin, M. A. A class of tests for multivariate normality based on linear functions of order statistics. Research Bulletin RB-72-16. Princeton, N.J.: Educational Testing Service, April, 1972.

Alexakos, C. E. Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions. Journal of Educational Psychology, 1966, 57, 297-306.

Alf, E. F. and Dorfman, D. D. The classification of individuals into two criterion groups on the basis of a discontinuous payoff function. Psychometrika, 1967, 32, 115-123.

Alumbaugh, R. V., Davis, H. G., and Sweney, A. B. A comparison of methods for constructing predictive instruments. Educational and Psychological Measurement, 1969, 29, 639-651.

Anderson, G. J., Walberg, H. J., and Welch, W. W. Curriculum effects on the social climate of learning: A new representation of discriminant functions. American Educational Research Journal, 1969, 6, 315-328.

Anderson, T. W. Introduction to multivariate statistical analysis. New York: Wiley, 1958.

Anderson, T. W. and Bahadur, R. R. Classification into two multivariate normal distributions with different covariance matrices. Annals of Mathematical Statistics, 1962, 33, 420-431.

Applebaum, M. and Bargmann, R. E. A FORTRAN II program for MUDAID: Multivariate, univariate, and discriminant analysis of irregular data. Technical Report. Urbana, Ill.: University of Illinois, 1967.

Asher, W. and Shively, J. E. The technique of discriminant analysis: A reclassification of Harbison and Myers' seventy-five countries. Comparative Education Review, 1969, 13, 180-186.

Baggaley, A. R., Isard, E. S., and Sherwood, E. J. Discrimination of academic curricula by the runner studies of attitude patterns - College Form. Measurement and Evaluation in Guidance, 1970, 3, 41-44.

Bargmann, R. E. Representative ordering and selection of variables. Final Report, Cooperative Research Project No. 1132, USOE, 1962a.

Bargmann, R. E. A method of classification based on dependent 0-1 patterns. Research Report, RC-677, IBM, 1962b.

Bargmann, R. E. Exploratory techniques involving artificial variables. In P. R. Krishnaiah (Ed.), Multivariate analysis II. New York: Academic Press, 1969. Pp. 567-580.

Bargmann, R. E. Interpretation and use of a generalized discriminant function. In R. C. Bose, et al. (Eds.), Essays in probability and statistics. Chapel Hill, N.C.: University of North Carolina Press, 1970. Pp. 35-60.

Bartlett, M. S. Multivariate analysis. Journal of the Royal Statistical Society, Series B, 1947, 9, 176-190.

Bartlett, M. S. and Please, N. W. Discrimination in the case of zero mean differences. Biometrika, 1963, 50, 17-21.

Bausell, R. B. and Magoon, J. Expected grade in a course, grade point average, and student ratings of the course and the instructor. Educational and Psychological Measurement, 1972, 32, 1013-1023.

Bledsoe, J. C. The prediction of teacher competence: A comparison of two multivariate statistical techniques. Multivariate Behavioral Research, 1973, 8, 3-22.

Bock, R. D. Contributions of multivariate experimental designs to educational research. In R. B. Cattell (Ed.), Handbook of multivariate experimental psychology. Chicago: Rand McNally, 1966. Pp. 820-840.

Bock, R. D. and Haggard, E. A. The use of multivariate analysis of variance in behavioral research. In D. K. Whitla (Ed.), Handbook of measurement and assessment in behavioral sciences. Reading, Mass.: Addison-Wesley, 1968. Pp. 100-143.

Bryan, J. G. Calibration of qualitative or quantitative variables for use in multiple-group discriminant analysis. Scientific Report No. 2 Contract No. AF19(604)-5207. Hartford, Conn.: The Travelers Insurance Companies, January, 1961.

Burnham, P. S. and Hewitt, B. A. Determining vocational characteristics versus predicting membership: A follow-up of male college graduates. Journal of Psychology, 1972, 81, 73-84.

Cacoullos, T. (Ed.) Discriminant analysis and applications. New York: Academic Press, 1973.

Chan, L. C. and Dunn, O. J. Treatment of missing values in discriminant analysis - I. The sampling experiment. Journal of the American Statistical Association, 1972, 67, 473-477.

Chapin, J. R. Patterns of characteristics of successful mathematics teachers and those who leave the profession: A case study. Mathematics Teacher, 1971, 63, 159-163.

Chastian, K. Prediction of success in audio-lingual and cognitive classes. Language Learning, 1969, 19, 27-39.

Cochran, W. G. On the performance of the linear discriminant function. Technometrics, 1964, 6, 179-190.

Cochran, W. G. and Hopkins, C. E. Some classification problems with multivariate qualitative data. Biometrics, 1961, 17, 10-32.

Cohen, D. Differentiating motivations underlying vocational choice. Journal of Educational Research, 1971, 64, 229-234.

Collier, R. O. A note on the multiple regression technique for deleting variables in the discriminant function. Journal of Experimental Education, 1963, 31, 351-353.

Cooley, W. W. and Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Cramer, E. M. Equivalence of two methods of computing discriminant function coefficients. Biometrics, 1967, 23, 153.

Cureton, E. E. Recipe for a cookbook. Psychological Bulletin, 1957, 54, 494-497.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Darlington, R. B., Weinberg, S. L., and Walberg, H. J. Canonical variate analysis and related techniques. Review of Educational Research, 1973, 43, 433-454.

Das Gupta, S. Theories and methods in classification: A review. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973. Pp. 77-137.

Dempster, A. P. Stepwise multivariate analysis of variance based on principal variables. Biometrics, 1963, 19, 478-490.

Desu, M. M. and Geisser, S. Methods and applications of equal-mean discrimination. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973. Pp. 139-159.

Dixon, W. J. (Ed.) Biomedical computer programs. Berkeley, Calif.: University of California Press, 1973.

Dixon, W. J. and Jenrich, R. I. Computer graphical analysis and discrimination. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973. Pp. 161-172.

Doerr, J. J. and Ferguson, J. L. The selection of vocational-technical students. Vocational Guidance Quarterly, 1968, 17, 27-32.

Draper, N. R. and Smith, H. Applied regression analysis. New York: Wiley, 1966.

Dunn, F. E. Two methods for predicting the selection of a college major. Journal of Counseling Psychology, 1959, 6, 15-27.

Dunn, O. J. Some expected values for probabilities of correct classification in discriminant analysis. Technometrics, 1971, 13, 345-353.

Dunsmore, I. R. A Bayesian approach to classification. Journal of the Royal Statistical Society, Series B, 1966, 28, 568-577.

Eisenbeis, R. A. and Avery, R. B. Discriminant analysis and classification procedures. Lexington, Mass.: Heath, 1972.

Eisenbeis, R. A., Gilbert, G. G., and Avery, R. B. Investigating the relative importance of individual variables and variable subsets in discriminant analysis. Communications in Statistics, 1973, 2, 205-219.

Elashoff, J. D., Elashoff, R. M., and Goldman, G. E. On the choice of variables in classification problems with dichotomous variables. Biometrika, 1967, 54, 668-670.

Field, T. W., et al. Identifying patterns of teacher behavior from student evaluations. Journal of Applied Psychology, 1971, 55, 469-469.

Fix, E. and Hodges, J. L. Discriminatory analysis, non-parametric discrimination: Consistency properties. Project No. 21-49-004. Report No. 4. Randolph Air Force Base, Texas: USAF School of Aviation Medicine, February, 1951.

Fix, E. and Hodges, J. L. Discriminatory analysis, non-parametric discrimination: Small sample performance. Project No. 21-49-004. Report No. 11. Randolph Air Force Base, Texas: USAF School of Aviation Medicine, August, 1952.

Friedman, H. P. and Rubin, J. On some invariant criteria for grouping data. Journal of the American Statistical Association, 1967, 62, 1159-1178.

Gabriel, K. R. Simultaneous test procedures in multivariate analysis of variance. Biometrika, 1968, 55, 489-504.

Geisser, S. Posterior odds for multivariate normal classifications. Journal of the Royal Statistical Society, Series B, 1964, 26, 69-76.

Geisser, S. Predictive discrimination. In P. R. Krishnaiah (Ed.), Multivariate analysis. New York: Academic Press, 1966. Pp. 149-163.

Geisser, S. Discriminatory practices. In D. L. Meyer and R. O. Collier (Eds.), Bayesian statistics. Itasca, Ill.: Peacock, 1970. Pp. 57-70.

Gessaman, M. P. and Gessaman, P. H. Comparison of some multivariate discrimination procedures. Journal of the American Statistical Association, 1972, 67, 468-472.

Gilbert, E. S. On discrimination using qualitative variables. Journal of the American Statistical Association, 1968, 63, 1399-1412.

Gilbert, E. S. The effect of unequal variance-covariance matrices on Fisher's linear discriminant function. Biometrics, 1969, 25, 505-515.

Glick, N. Sample-based classification procedures derived from density estimators. Journal of the American Statistical Association, 1972, 67, 116-122.

Goldman, R. D. and Warren, R. Discriminant analysis of study strategies connected with college grade success in different major fields. Journal of Educational Measurement, 1973, 10, 39-47.

Gregson, R. A. M. Sorting sheep from goats with one yardstick. Occupational Psychology, 1964, 38, 191-201.

Grimsley, G. and Summers, G. W. Selection techniques for Pakistani postgraduate students of business. Educational and Psychological Measurement, 1965, 25, 1133-1142.

Grizzle, J. E. An example of the analysis of a series of response curves and an application of multivariate multiple comparisons. In R. C. Bose, et al. (Eds.), Essays in probability and statistics. Chapel Hill, N.C.: University of North Carolina Press, 1970. Pp. 311-326.

Hall, C. E. Three papers in multivariate analysis. Palo Alto, Calif.: American Institute for Research, 1967.

Hall, C. E. Rotation of canonical variates in multivariate analysis of variance. Journal of Experimental Education, 1969, 38, 31-38.

Harman, H. H. How factor analysis can be used in classification. Research Bulletin, RB-71-65. Princeton, N.J.: Educational Testing Service, 1971.

Harris, C. W. Note on dispersion analysis. Journal of Experimental Education, 1954, 22, 289-291.

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart and Winston, 1973.

Healy, M. J. R. Computing a discriminant function from within-sample dispersions. Biometrics, 1969, 21, 1011-1012.

Henschke, C. I., Kossack, C. F., and Lissitz, R. W. A generalized classification technique using interval, ordinal, and nominal variables. Multivariate Behavioral Research, 1974, 9, 105-118.

Hills, M. Allocation rules and their error rates. Journal of the Royal Statistical Society, Series B, 1966, 28, 1-31.

Hills, M. Discrimination and allocation with discrete data. Applied Statistics, 1967, 16, 237-250.

Hockersmith, F. B. Evaluation of new approaches to classification error rates. Special Report, Project No. 6-2151, Grant No. OEG-1-7-062151-0571, HEW, 1969.

Hodges, J. L. Discriminatory analysis. I. Survey of discriminatory analysis. Report No. 1, Project No. 21-49-004. Randolph Air Force Base, Texas: USAF School of Aviation Medicine, October, 1950.

Horst, P. The prediction of personal adjustment. New York: Social Science Research Council, 1941.

Horst, P. Multiple classification by the method of least squares. Journal of Clinical Psychology, 1956a, 12, 13-16.

Horst, P. Least square multiple classification for unequal subgroups. Journal of Clinical Psychology, 1956b, 12, 309-315.

Horst, P. Factor analysis of data matrices. New York: Holt, Rinehart and Winston, 1965.

Horst, P. An overview of the essentials of multivariate analysis methods. In R. B. Cattell (Ed.), Handbook of multivariate experimental psychology. Chicago: Rand McNally, 1966. Pp. 129-152.

Horton, I. R., Russell, J. S., and Moore, A. W. Multivariate-covariance and canonical analysis: A method for selecting the most effective discriminators in a multivariate situation. Biometrics, 1968, 24, 845-858.

Huberty, C. J An empirical comparison of selected classification rules in multiple group discriminant analysis. Unpublished doctoral dissertation, University of Iowa, 1969.

Huberty, C. J On the variable selection problem in multiple group discriminant analysis. Paper presented at the Annual Meeting of the American Educational Research Association, New York, February, 1971a.

Huberty, C. J The equivalence of two multivariate classification schemes. Multivariate Behavioral Research, 1971b, 6, 243-246.

Huberty, C. J Multivariate indices of strength of association. Multivariate Behavioral Research, 1972a, 7, 523-526.

Huberty, C. J Regression analysis and 2-group discriminant analysis. Journal of Experimental Education, 1972b, 41, 39-41.

Huberty, C. J Use of some "discriminant analysis" computer programs. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1974a.

Huberty, C. J Covariance and discriminant analysis. Paper presented at the Annual Meeting of the Southeastern Psychological Association, Hollywood, Florida, May, 1974b.

Huberty, C. J and Blommers, P. J. An empirical comparison of three indices of variable contribution in multiple group discriminant analysis. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1972.

Huberty, C. J and Blommers, P. J. An empirical comparison of the accuracy of selected multivariate classification rules. Multivariate Behavioral Research, 1974, 9, 59-84.

Ito, K. On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures. In P. R. Krishnaiah (Ed.), Multivariate analysis II. New York: Academic Press, 1969. Pp. 87-120.

Jackson, E. C. Missing values in linear multiple discriminant analysis. Biometrics, 1968, 24, 835-844.

James, G. S. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. Biometrika, 1954, 41, 19-43.

Jones, L. V. Analysis of variance in its multivariate developments. In R. B. Cattell (Ed.), Handbook of multivariate experimental psychology. Chicago: Rand McNally, 1966. Pp. 244-266.

Keenen, C. B. and Holmes, J. E. Predicting graduation withdrawal and failure in college by multiple discriminant analysis. Journal of Educational Measurement, 1970, 7, 91-95.

Kendall, M. G. Discriminantion and classification. In P. R. Krishnaiah (Ed.), Multivariate analysis. New York: Academic Press, 1966. Pp. 165-185.

Kendall, M. G. A course in multivariate analysis. New York: Hafner, 1968.

Kendall, M. G. The basic problems of cluster analysis. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973. Pp. 179-191.

Kirkendall, D. R. and Ismail, A. H. The ability of personality variables in discriminating among three intellectual groups of preadolescent boys and girls. Child Development, 1970, 41,1173-1181.

Knutsen, M. An empirical comparison of the linear discriminant function and multiple regression techniques in the classifying of subjects into three categories. Unpublished doctoral dissertation, University of Washington, 1955.

Kossack, C. F. On the mechanics of classification. Annals of Mathematical Statistics, 1945, 16, 95-98.

Kossack, C. F. A generalized classification statistic. Technical Report No. 16. Contract No. AF41(609)-3145. San Antonio, Texas: Personnel Research Laboratory, Lackland Air Force Base, May, 1967a.

Kossack, C. F. Non-parametric statistical classification. Technical Report No. 17. Contract No. AF41(609)-3145. San Antonio, Texas: Personnel Research Laboratory, Lackland Air Force Base, June, 1967b.

Kruglick, F. and Brewer, J. K. Application of a discrete multivariate procedure for classifying criterion referenced test outcomes. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, April, 1974.

Kshirsagar, A. M. Multivariate analysis. New York: Dekker, 1972.

Kundert, K. R. Tools of analysis for pattern recognition. Unpublished doctoral dissertation, University of Georgia, 1972.

Lachenbruch, P. A. Discriminant analysis when the initial samples are misclassified. Technometrics, 1966, 8, 657-662.

Lachenbruch, P. A. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 1967, 23, 639-645.

Lachenbruch, P. A. On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. Biometrics, 1968, 24, 823-834.

Lachenbruch, P. A. Some results on the multiple group discrimination problem. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973. Pp. 193-211.

Lachenbruch, P. A. and Mickey, M. R. Estimation of error rates in discriminant analysis. Technometrics, 1968, 10, 1-11.

Lachenbruch, P. A., Sneeringer, C., and Revo, L. T. Robustness of the linear and quadratic discriminant functions to certain types of non-normality. Communications in Statistics, 1973, 1, 39-56.

Lissitz, R. W. and Henschke-Mason, C. The selection of independent variables and prior probabilities as a factor influencing the accuracy of classifying individuals into existing groups. Multivariate Behavioral Research, 1972, 7, 489-497.

Lissitz, R. W. and Schoenfeldt, L. F. Moderator subgroups for the estimation of educational performance: A comparison of prediction models. American Educational Research Journal, 1974, 11, 63-75.

Lockhart, R. S. The assumption of multivariate normality. British Journal of Mathematical and Statistical Psychology, 1967, 20, 63-69.

Lohnes, P. R. Test space and discriminant space classification models and related significance tests. Educational and Psychological Measurement, 1961, 21, 559-574.

Lohnes, P. R. and Gribbons, W. D. The Markov chain as a null hypothesis in a development survey. Journal of Educational Measurement, 1970, 7, 25-32.

Lubin, A. Linear and non-linear discriminating functions. British Journal of Mathematical and Statistical Psychology, 1950, 3, 90-104.

Malkovich, J. F. and Afifi, A. A. On tests for multivariate normality. Journal of the American Statistical Association, 1973, 68, 176-179.

Mantel, N. Why stepdown procedures in variable selection. Technometrics, 1970, 12, 621-625.

Maw, W. H. and Magoon, A. J. The curiosity dimension of fifth-grade children: A factorial discriminant analysis. Child Development, 1971, 42, 2023-2031.

McCall, R. B. The use of multivariate procedures in developmental psychology. In P. H. Mussen (Ed.), Manual of Child Psychology. New York: Wiley, 1970. Pp. 1366-1377.

McDonald, L. On the estimation of missing data in the multivariate linear model. Biometrics, 1971, 27, 535-543.

McHugh, R. B. and Apostolakos, P. C. Methodology for the comparions of clinical with actuarial predictions. Psychological Bulletin, 1959, 56, 301-308.

McNeil, K. A. Semantic space as an indicator of socialization. Journal of Educational Psychology, 1968, 59, 325-327.

McQuarrie, D. and Grotelueschen, A. Effects of verbal warning upon misapplication of a rule of limited applicability. Journal of Educational Psychology, 1971, 62, 432-438.

Meehl, P. E. and Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. Psychological Bulletin, 1955, 52, 194-216.

Michael, W. B. and Perry, N. C. The comparability of the single discriminant function and multiple regression techniques. Journal of Experimental Education, 1956, 24, 297-201.

Michaelis, J. Simulation experiments with multiple group linear and quadratic discriminant analysis. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press, 1973. Pp. 225-238.

Miller, R. G. Selecting predictors for multiple discriminant analysis. Meteorological Monographs, 1962, 4, 11-14.

Mises, R. von. On the classification of observation data into distinct groups. Annals of Mathematical Statistics, 1945, 16, 68-73.

Molnar, G. E. and Delauretis, R. J. Predicting the curriculum mobility of engineering students: A comparison of discriminant procedures. Journal of Counseling Psychology, 1973, 20, 50-59.

Moore, D. H. Evaluation of five discrimination procedures for binary variables. Journal of the American Statistical Association, 1973, 68, 399-404.

Mulaik, S. A. The foundations of factor analysis. New York: McGraw-Hill, 1972.

Neal, R. and King, P. Comparison of a multivariate and a configural analysis for classifying engineering students. Journal of Counseling Psychology, 1969, 16, 563-568.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Ottman, D. K., Ferguson, J. L., and Kaufman, M. B. A discriminatory study for classification of Navy AN(P) school graduates into Navy class "A" school assignments. Technical Report No. 7 for Contract No. NONR 649(00) between University of Missouri and Office of Naval Research, October, 1956.

Overall, J. E. and Klett, C. J. Applied multivariate analysis. New York: McGraw-Hill, 1972.

Porebski, O. R. On the interrelated nature of the multivariate statistics used in discriminatory analysis. British Journal of Mathematical and Statistical Psychology, 1966a, 19, 197-214.

Porebski, O. R. Discriminatory and canonical analysis of technical college data. British Journal of Mathematical and Statistical Psychology, 1966b, 19, 215-236.

Press, S. J. Applied multivariate analysis. New York: Holt, Rinehart and Winston, 1972.

Pruzek, R. M. Methods and problems in the analysis of multivariate data. Review of Educational Research, 1971, 41, 163-190.

Raiffa, H. Statistical decision theory approach to item selection for dichotomous test and criterion variables. In H. Solomon (Ed.), Studies in item analysis and prediction. Stanford, Calif.: Stanford University Press, 1961. Pp. 187-220.

Rao, C. R. On some problems arising out of discrimination with multiple characters. Sankhyā, 1949, 9, 343-366.

Rao, C. R. Advanced statistical methods in biometric research. New York: Wiley, 1952.

Rao, C. R. Linear statistical inference and its applications. New York: Wiley, 1965.

Rock, D. A., Baird, L. L., and Linn, R. L. Interaction between college effects and students aptitudes. American Educational Research Journal, 1972, 9, 149-161.

Rogers, G. and Linden, J. D. Use of multiple discriminant function analysis in the evaluation of three multivariate grouping techniques. Educational and Psychological Measurement, 1973, 33, 787-802.

Roy, S. N. and Bargmann, R. E. Tests of multiple independence and the associated confidence bounds. Annals of Mathematical Statistics, 1958, 29, 491-503.

Rozeboom, W. W. Foundations of the theory of prediction. Homewood, Ill.: Dorsey, 1966.

Rubin, D. B. Missing at random - what does it mean? Research Bulletin, RB-73-2. Princeton, N.J.: Educational Testing Service, January, 1973.

Rulon, P. J., et al. Multivariate statistics for personnel classification. New York: Wiley, 1967.

Rulon, P. J. and Brooks, W. D. On statistical tests of group differences. In D. K. Whitla (Ed.), Handbook of measurement and assessment in behavioral sciences. Reading, Mass.: Addison-Wesley, 1968. Pp. 60-99.

Saupe, J. L. Factorial-design multiple discriminant analysis: A description and an illustration. American Educational Research Journal, 1965, 2, 175-184.

Schoenfeldt, L. F. Education after high school. Sociology of Education, 1968, 41, 350-369.

Schoenfeldt, L. F. Nontest variables in the prediction of educational progress. Paper presented at the Annual Meeting of the American Psychological Association, Miami, September, 1970.

Smith, A. C., et al. Prediction of developmental outcome at seven years from prenatal, perinatal, and postnatal events. Child Development, 1972, 43, 495-507.

Smith, C. A. B. Some examples of discrimination. Annals of Eugenics, 1947, 13, 272-282.

Smith, W. B. and Zeis, C. D. On classification for incomplete multinomial data. Communications in Statistics, 1973, 2, 85-93.

Solomon, H. Classification procedures based on dichotomous response vectors. In H. Solomon (Ed.), Studies in item analysis and prediction. Stanford, Calif.: Stanford University Press, 1961. Pp. 177-186.

Sorum, M. J. Estimating the conditional probability of misclassification. Technometrics, 1971, 13, 333-343.

Sorum, M. J. Three probabilities of misclassification. Technometrics, 1972a, 14, 309-316.

Sorum, M. J. Estimating the expected and the optimal probabilities of misclassification. Technometrics, 1972b, 14, 935-943.

Southworth, J. A. and Morningstar, M. E. Persistence of occupational choice and personality congruence. Journal of Counseling Psychology, 1970, 17, 409-412.

Spain, W. H. and D'Costa, A. G. Estimating the significance of multivariate experimental effects in the presence of confounding interactions. Paper presented at the Annual Meeting of the American Educational Research Association, Minneapolis, March, 1970.

Stahmann, R. F. Predicting graduation major field from freshman entrance data. Journal of Counseling Psychology, 1969, 16, 109-113.

Stevens, J. P. Global measures of association in multivariate analysis of variance. Multivariate Behavioral Research, 1972a, 7, 373-378.

Stevens, J. P. Four methods of analyzing between variation for the k-group MANOVA problem. Multivariate Behavioral Research, 1972b, 7, 499-522.

Stevens, J. P. Step-down analysis and simultaneous confidence intervals in MANOVA. Multivariate Behavioral Research, 1973, 8, 391-402.

Stevens, J. P. and Barcikowski, R. S. A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1974.

Tatsuoka, M. M. Joint probability of membership in a group and success therein: An index which combines the information from discriminant and regression analyses. Unpublished doctoral dissertation, Harvard University, 1956.

Tatsuoka, M. M. Multivariate analysis. Review of Educational Research, 1969, 39, 739-743.

Tatsuoka, M. M. Discriminant analysis. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.

Tatsuoka, M. M. Multivariate analysis. New York: Wiley, 1971.

Tatsuoka, M. M. Multivariate analysis in educational research. In F. N. Kerlinger (Ed.), Review of research in education. Itasca, Ill.: Peacock, 1973a. Pp. 273-319.

Tatsuoka, M. M. An examination of the statistical properties of a multivariate measure of strength of relationship. Final Report, Project No. 2-E-020, Grant No. OEG-5-72-0027(509), HEW, December, 1973b.

Tatsuoka, M. M. and Tiedeman, D. V. Discriminant analysis. Review of Educational Research, 1954, 24, 402-420.

Thorndike, R. M. and Weiss, D. J. A study of the stability of canonical correlations and canonical components. Educational and Psychological Measurement, 1973, 33, 123-134.

Tiedeman, D. V. The utility of the discriminant function in psychological and guidance investigations. Harvard Educational Review, 1951, 21, 71-95.

Timm, N. H. The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 1970, 35, 417-437.

Timm, N. H. Multivariate analysis with applications in education and psychology. Belmont, Calif.: Brooks/Cole, 1974.

Tintner, G. Some formal relations in multivariate analysis. Journal of the Royal Statistical Society, Series B, 1950, 12, 95-101.

Toms, J. G. and Brewer, J. K. A discrete discriminant analysis procedure for vocational rehabilitation predictions. Rehabilitation Literature, 1971, 32, 232-238.

Urbakh, V. Y. Linear discriminant analysis: Loss of discriminating power when a variate is omitted. Biometrics, 1971, 27, 531-534.

Van de Geer, J. P. Introduction to multivariate analysis for the social sciences. San Francisco: Freeman, 1971.

Veldman, D. J. FORTRAN programming for the behavioral sciences. New York: Holt, Rinehart and Winston, 1967.

Ward, J. H. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 1963, 58, 236-244.

Weiner, J. M. and Dunn, O. J. Elimination of variates in linear discrimination problems. Biometrics, 1966, 22, 268-275.

Welch, B. L. Note on discriminant functions. Biometrika, 1939, 31, 218-220.

Whellams, F. S. Musical abilities and sex differences in the analysis of aural-musical capacities. Journal of Research in Music Education, 1973, 21, 30-39.

Williams, W. S. A study of the use of the semantic differential by fifth grade children from different socioeconomic groups. Journal of Psychology, 1972, 81, 343-350.

Wood, D. A. Background characteristics and work values distinguishing satisfaction levels among engineers. Journal of Applied Psychology, 1971, 55, 537-542.

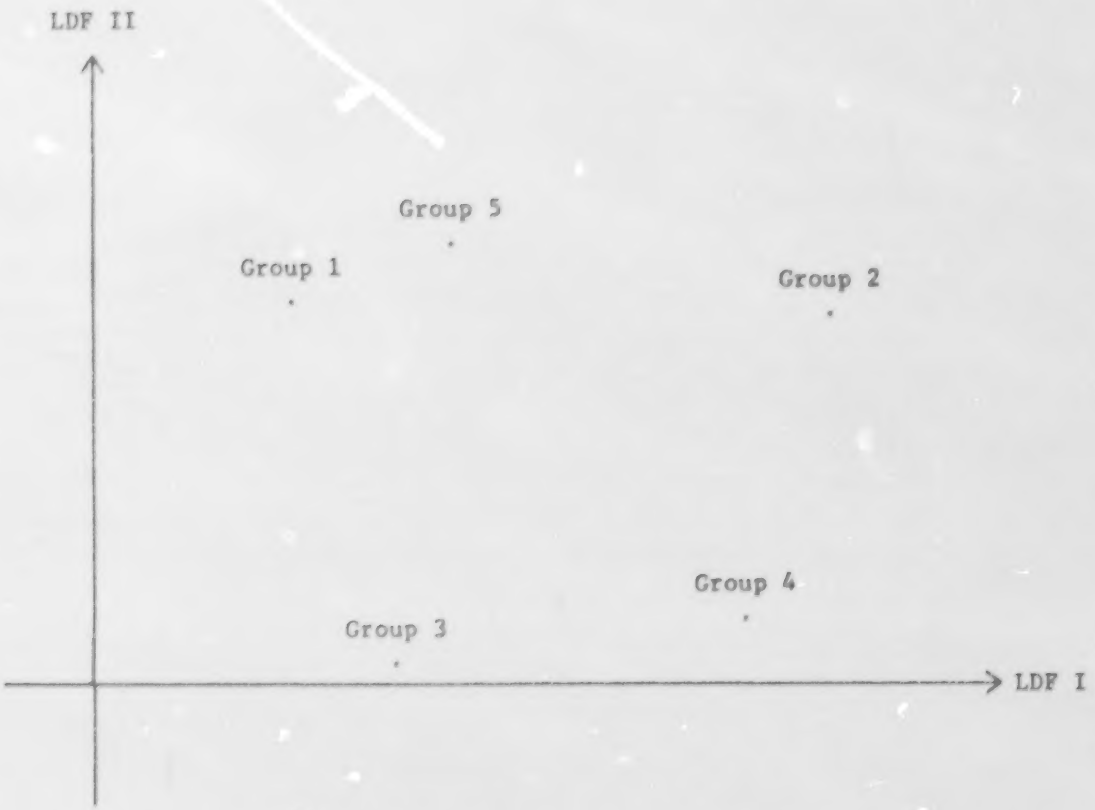

Figure 1. Hypothetical Centroids in Discriminant Plane

Table 1

Classification Statistics

Linear

[13] $$D_{ik} = [(\underline{X}_i - \overline{\underline{X}}_k)' S^{-1} (\underline{X}_i - \overline{\underline{X}}_k)]^{\frac{1}{2}}$$

[14] $$L_{ik} = -\tfrac{1}{2} \ln|S| - \tfrac{1}{2} D_{ik}^2 + \ln p_k$$

[15] $$P_{ik} = \frac{p_k \exp(-\frac{1}{2} D_{ik}^2)}{\sum_{k'=1}^{K} p_{k'} \exp(-\frac{1}{2} D_{ik'}^2)}$$

[16] $$Q_{ik} = \frac{p_k \cdot h_{ik}}{\sum_{k'=1}^{K} p_{k'} \cdot h_{ik'}}$$

Quadratic

[17] $$D'_{ik} = [(\underline{X}_i - \overline{\underline{X}}_k)' S_k^{-1} (\underline{X}_i - \overline{\underline{X}}_k)]^{\frac{1}{2}}$$

[18] $$L'_{ik} = -\tfrac{1}{2} \ln|S_k| - \tfrac{1}{2}(D'_{ik})^2 + \ln p_k$$

[19] $$P'_{ik} = \frac{p_k |S_k|^{-\frac{1}{2}} \exp[-\frac{1}{2}(D'_{ik})^2]}{\sum_{k'=1}^{K} p_{k'} |S_{k'}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(D'_{ik'})]^2}$$

[20] $$Q'_{ik} = \frac{p_k \cdot g_{ik}}{\sum_{k'=1}^{K} p_{k'} \cdot g_{ik'}}$$

Regression

[21] $$Y_{ik} = \underline{b}'_k (\underline{X}_i - \overline{\underline{X}}) + \overline{Y}_k$$